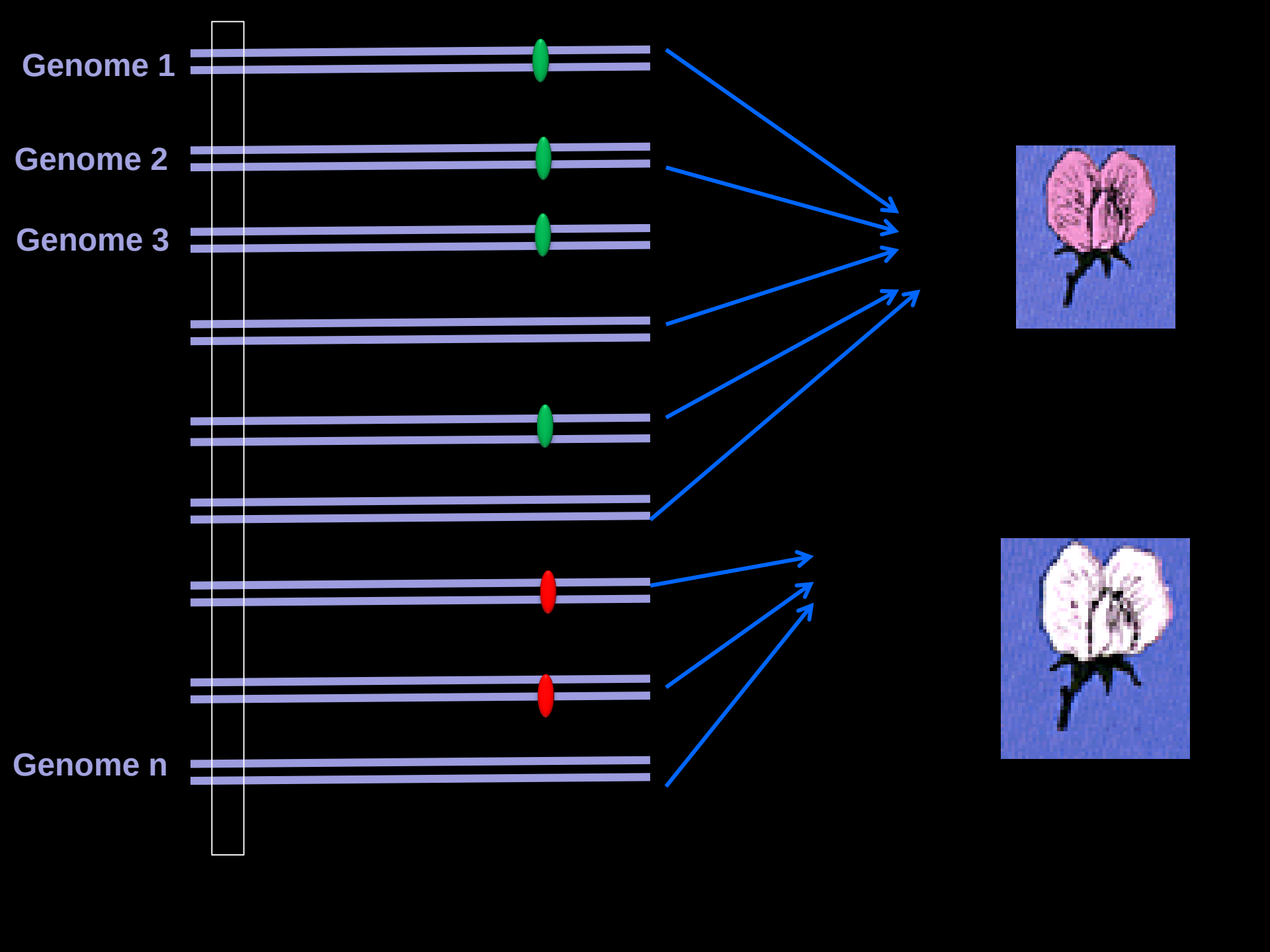
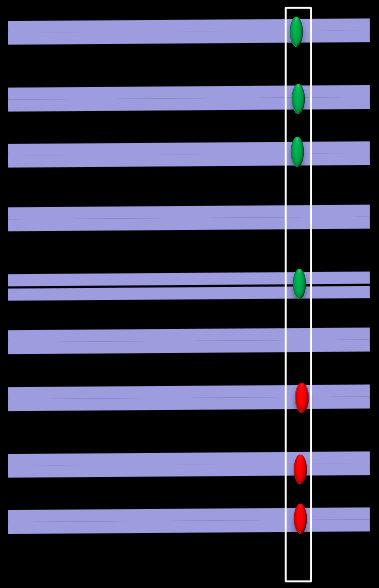
Towards bridging the genotype-phenotype gap

systems level dissection of the impact of variation in DNA sequence and structure on gene expression

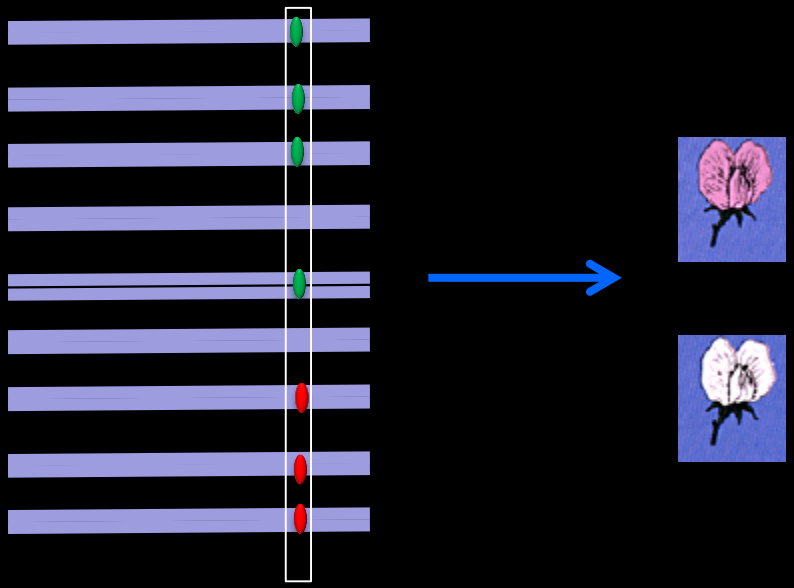
Teresa Przytycka

NIH / NLM / NCBI

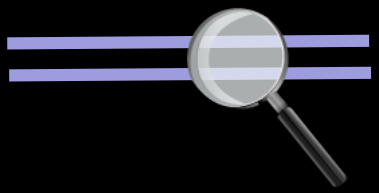

Genome 1
Genome 2
Genome 3
Genome n



- association

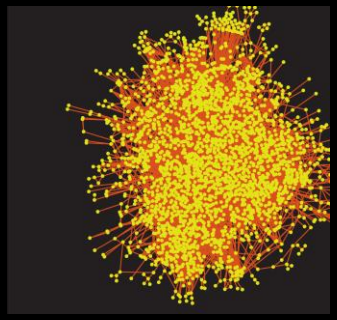


- association
- specific gene / mutation
- Mechanism of action
- Effect on other molecules/pathways?
- Complexity of phenotypes
- epistatic effects



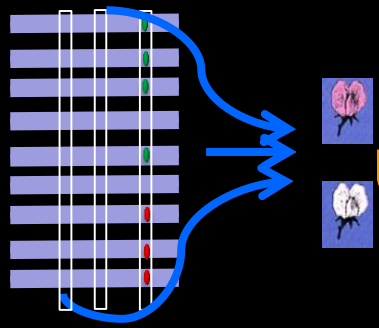
Impact of variations in DNA/RNA sequence and structure on gene expression and function

- Conformational changes in DNA structure
- Conformational changes in RNA structure



Propagation of genetic perturbations through the molecular system

- Propagation of the effect of copy number variations
- Identification of pathways dys-regulated in cancer
- Identification of “causal” genes



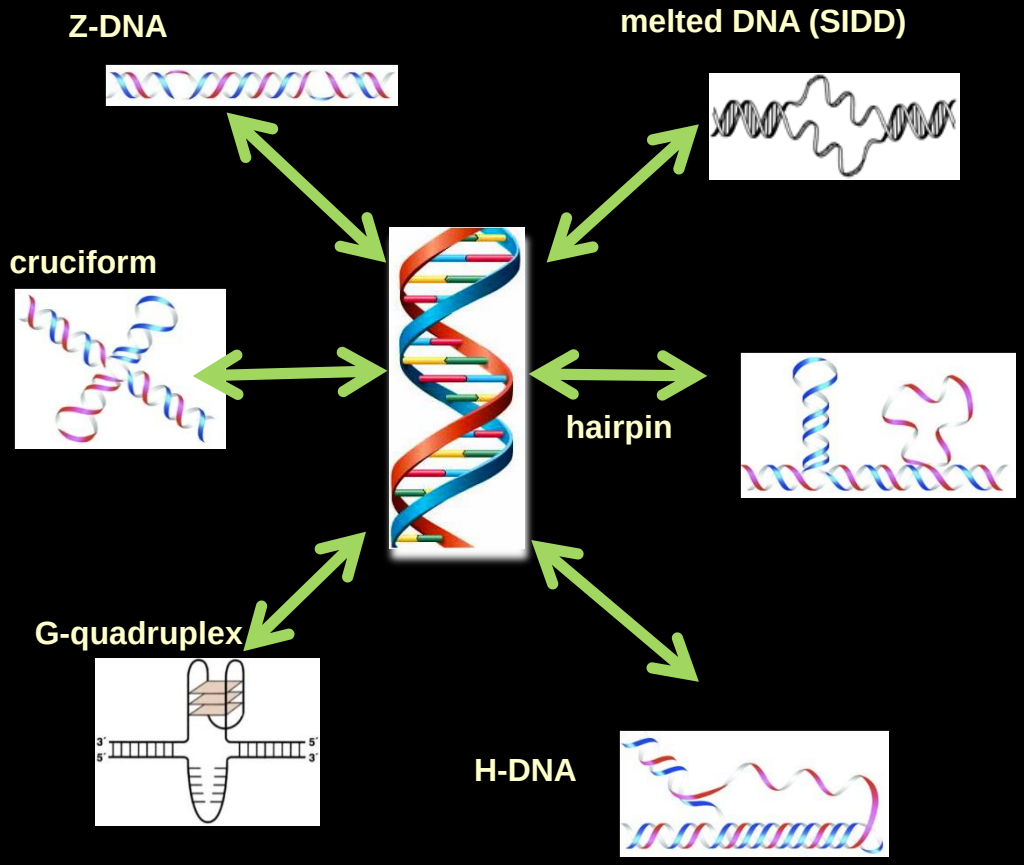
Uncovering epsitatic interactions

- Identification of epistatic interactions involved in drug resistance

Conformational changes of DNA structure

collaboration with *David Levens* (NCI) and *Rafael Casellas* (NIAMS)

Special acknowledgments:
Damian Wojtowicz, Fedor Kouzine, Arito Yamane

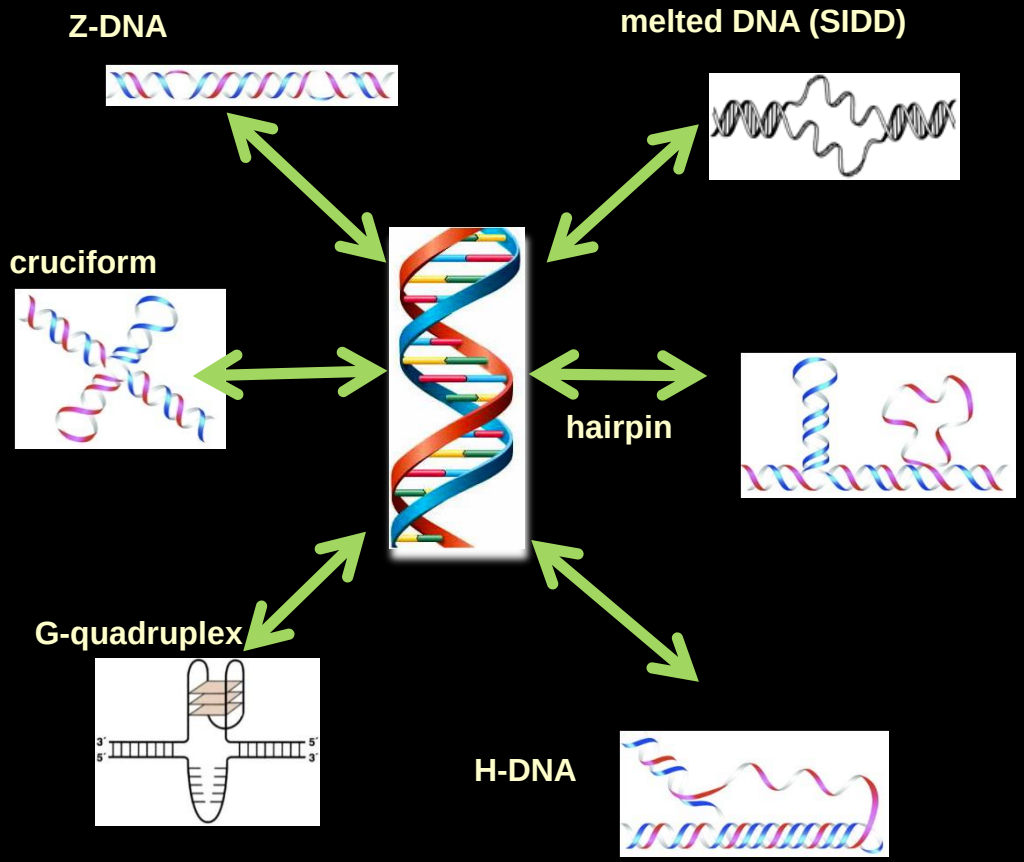


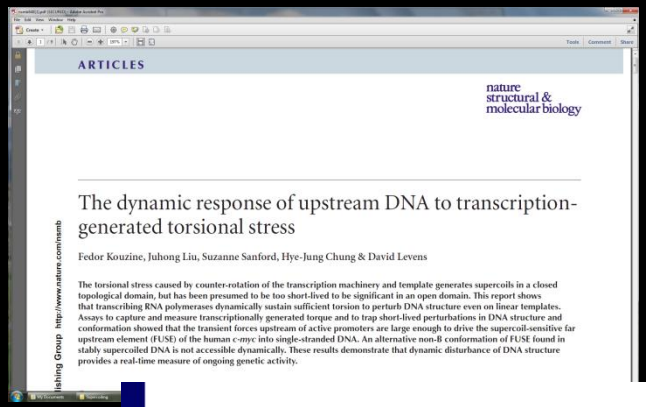
Conformational changes in DNA structure

collaboration with *David Levens* (NCI) and *Rafael Casellas* (NIAMS)

Special acknowledgments:

Damian Wojtowicz (TP), *Fedor Kouzine* (DL), *Arito Yamada* (RC)




ARTICLES

nature structural & molecular biology

The dynamic response of upstream DNA to transcription-generated torsional stress

Fedor Kouzine, Juhong Liu, Suzanne Sanford, Hye-Jung Chung & David Levens

The torsional stress caused by counter-rotation of the transcription machinery and template generates supercoils in a closed topological domain, but has been presumed to be too short-lived to be significant in an open domain. This report shows that transcribing RNA polymerases dynamically sustain sufficient torsion to perturb DNA structure even on linear templates. Assays to capture and measure transcriptionally generated torque and to trap short-lived perturbations in DNA structure and conformation showed that the transient forces upstream of active promoters are large enough to drive the supercoil-sensitive far upstream element (FUSE) of the human *c-myc* into single-stranded DNA. An alternative non-B conformation of FUSE found in stably supercoiled DNA is not accessible dynamically. These results demonstrate that dynamic disturbance of DNA structure provides a real-time measure of ongoing genetic activity.

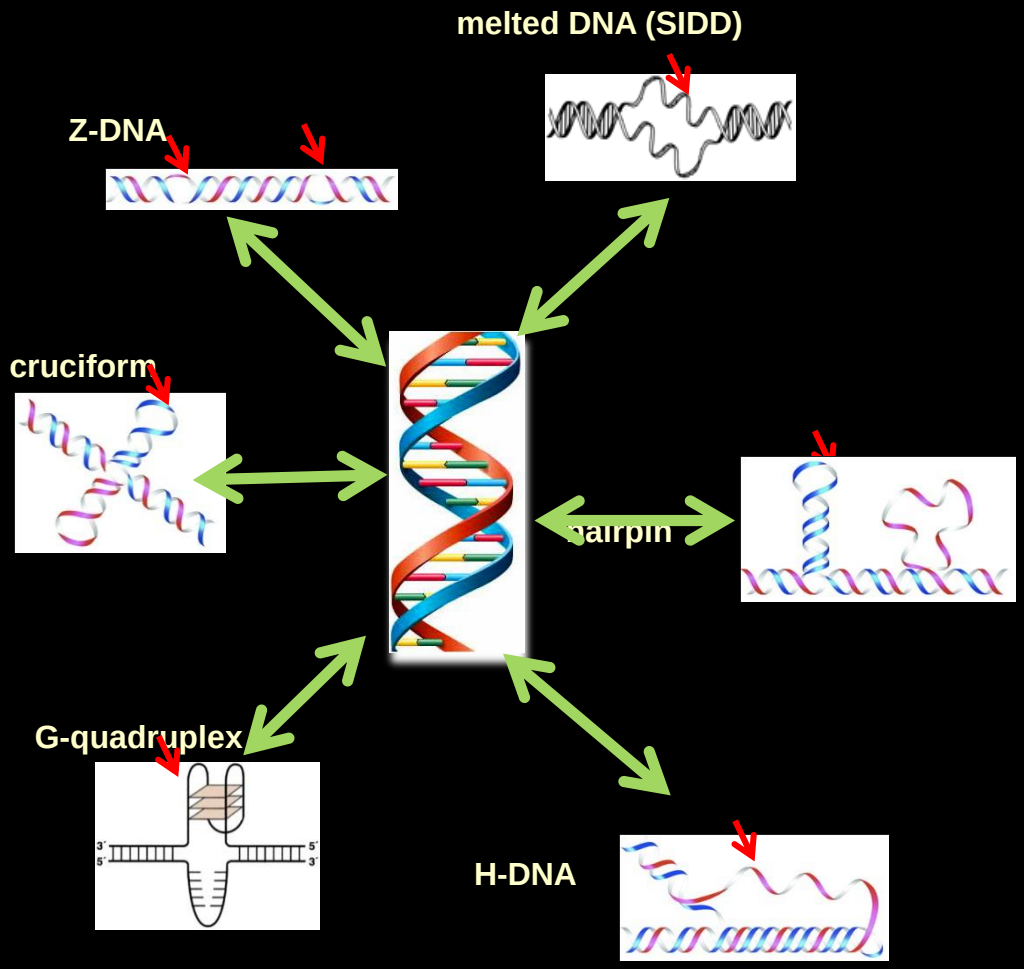
Z-DNA-forming silencer in the first exon reg
human ADAM-12 gene expression

Bimal K. Ray, Srijita Dhar, Arvind Shakya, and Alpana Ray[1]

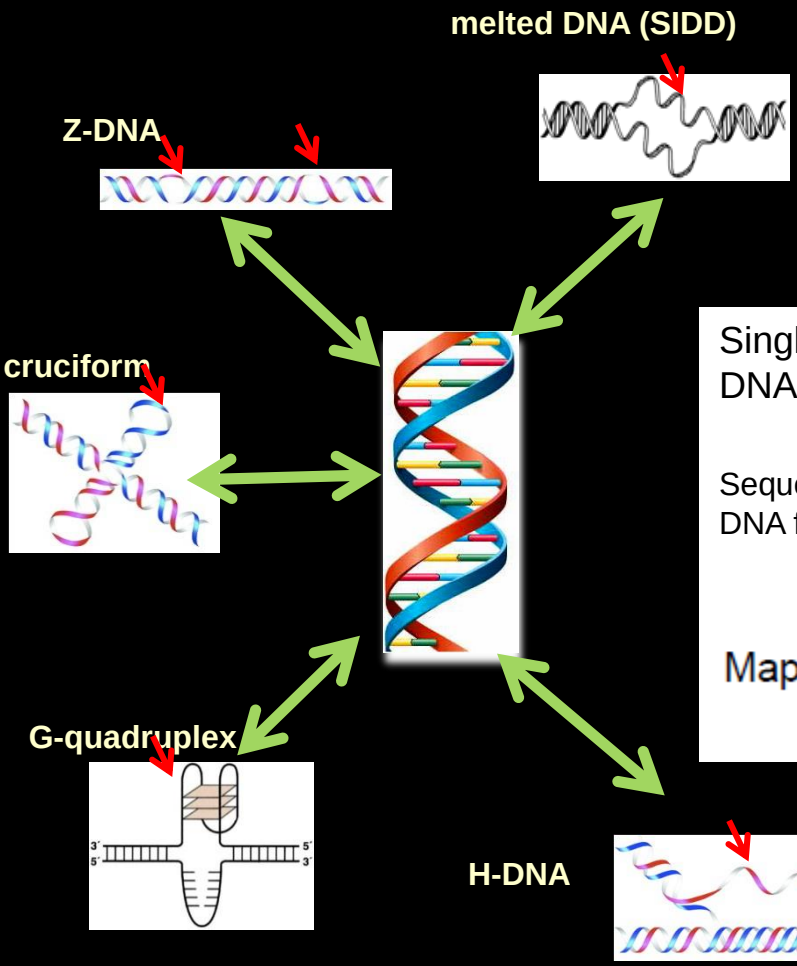
Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211

Edited* by Alexander Rich, Massachusetts Institute of Technology, Cambridge, MA, and approved November 19, 2010 (receiv

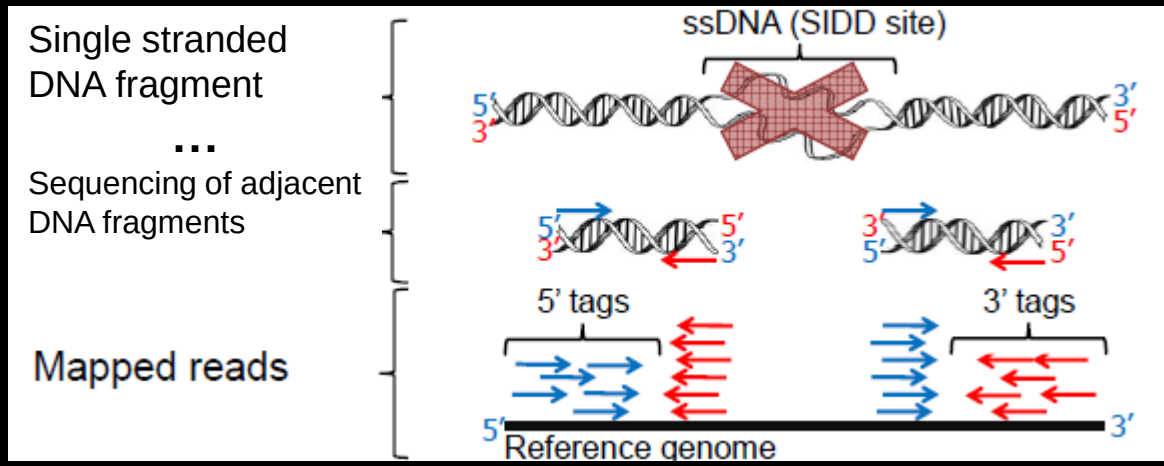
Detecting Non-B DNA structures in *vivo*



Detecting Non-B DNA structures in *vivo*



Detecting unpaired bases: ssDNA - Seq



Detecting Non-B DNA structures in *vivo*

A+T rich sequence

melted DNA (SIDD)

$(YR{\cdot}YR)_n$

Z-DNA

cruciform

inverted repeats

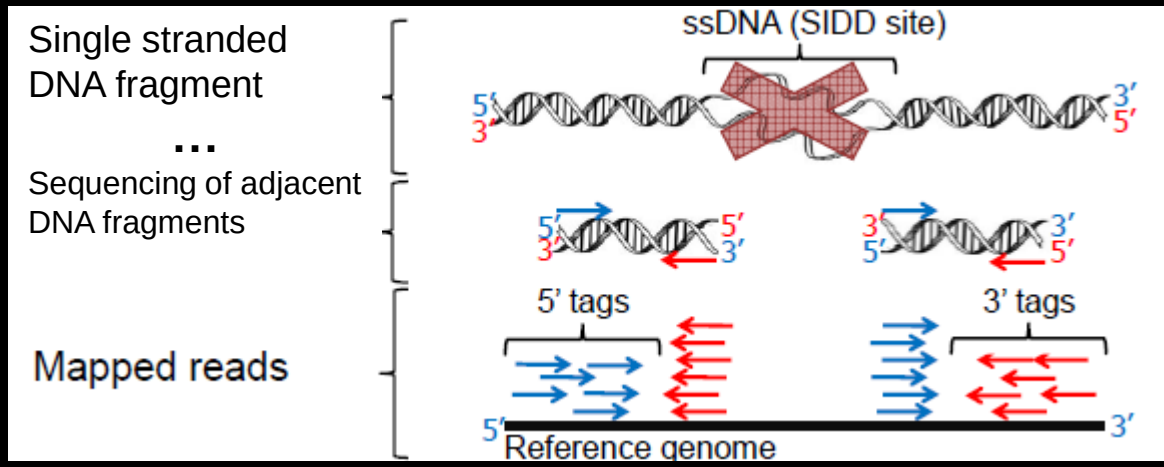
G-quadruplex

$(G)_n$ tracts

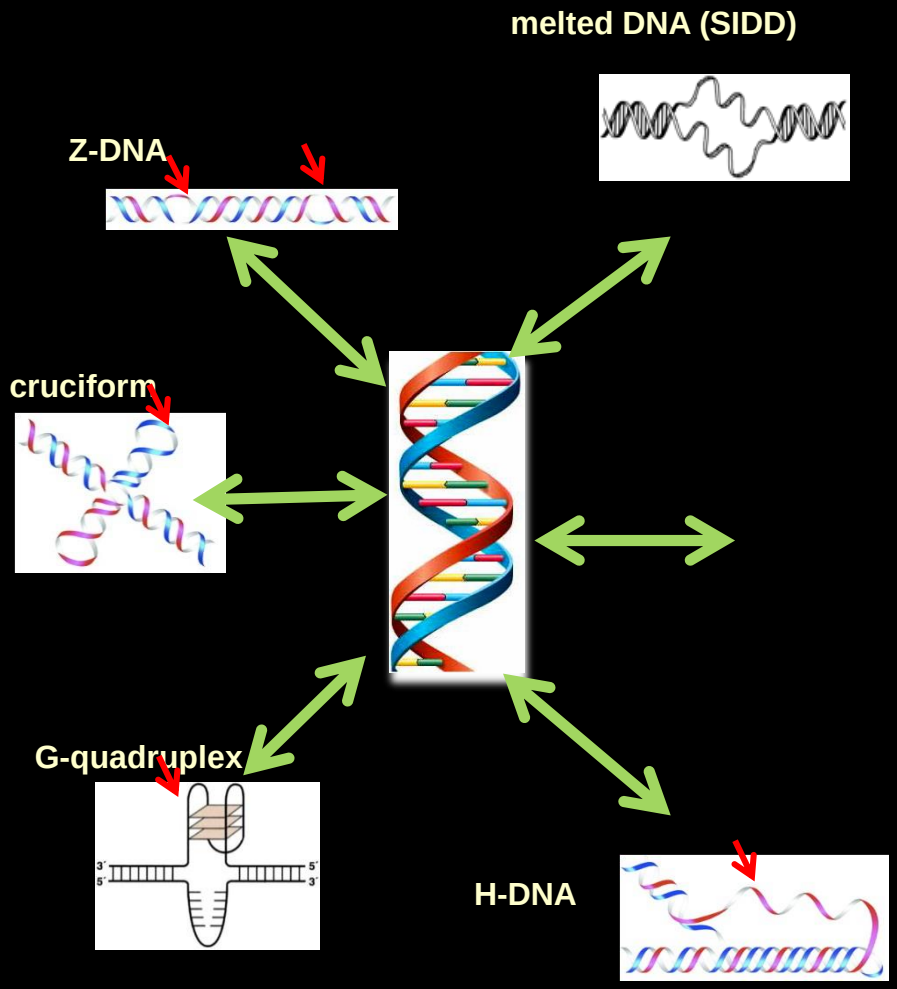
$(Y{\cdot}R)_n$ mirror repeats

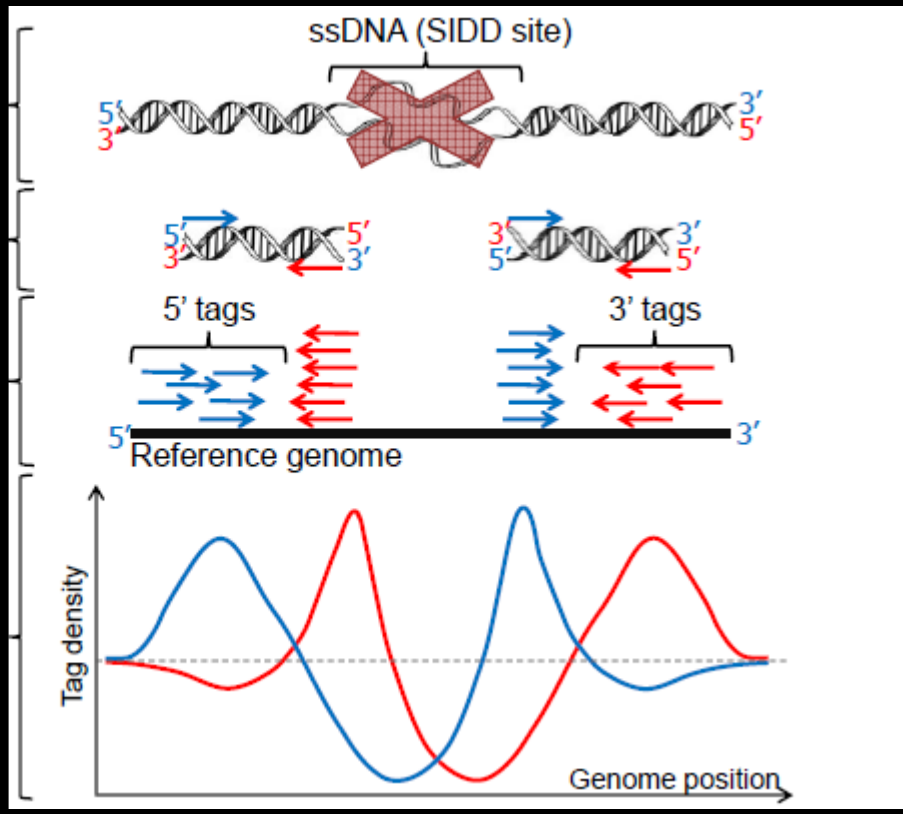
H-DNA

Detecting unpaired bases: ssDNA - Seq



Specific non-B DNA structures are accompanied by specific tag distribution

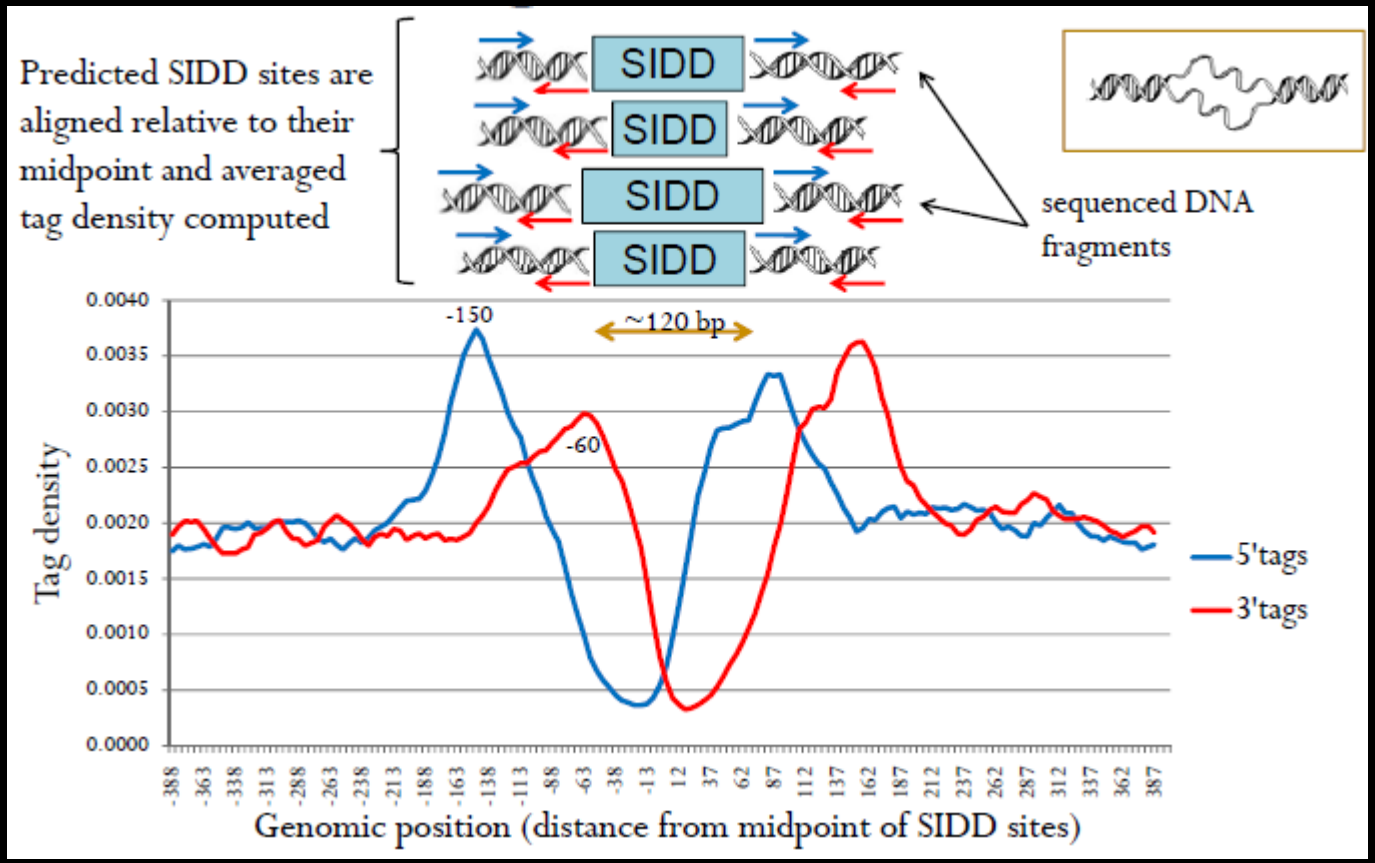




Expected tags distribution for SIDD

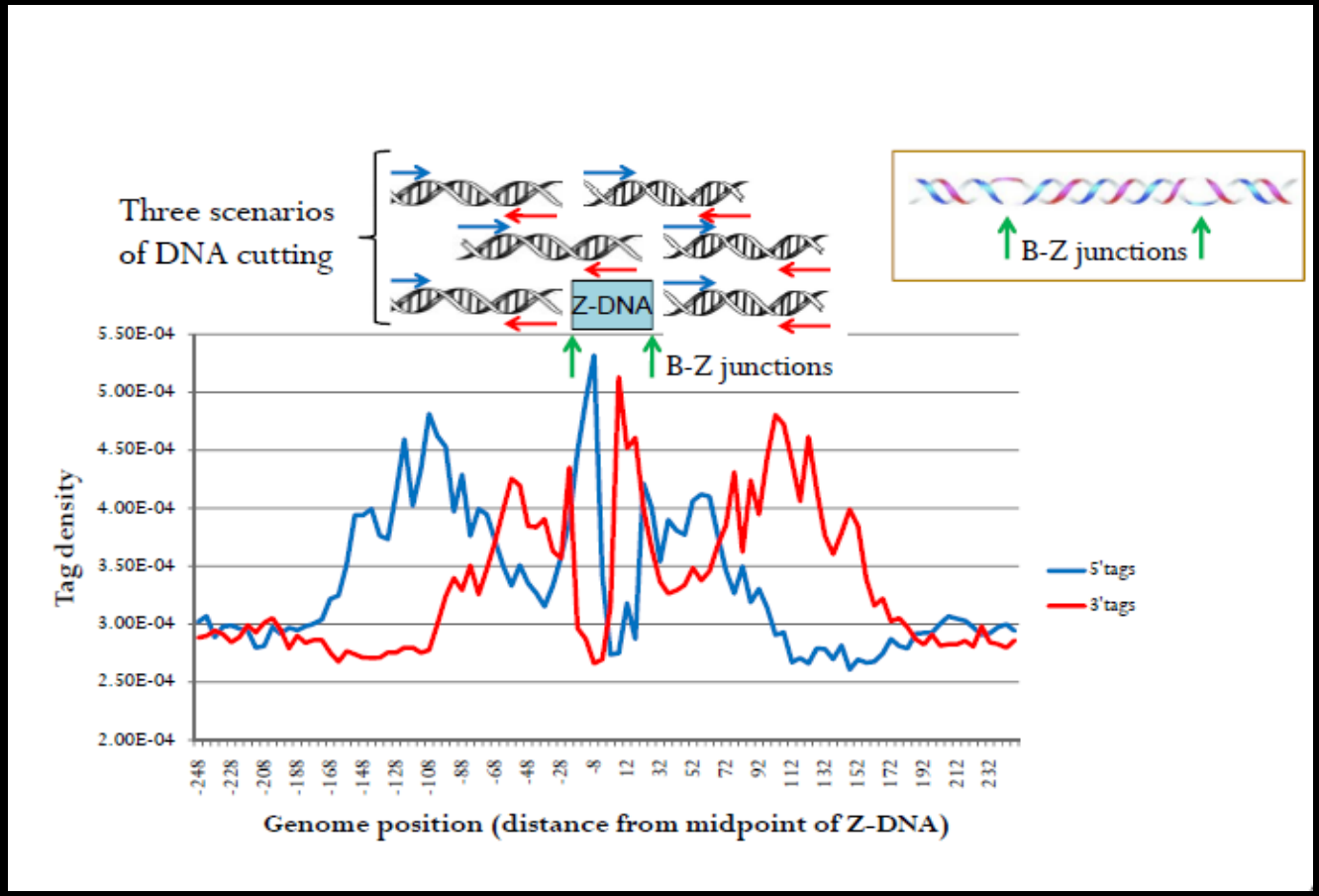
Specific non-B DNA structures are accompanied by specific tag distribution

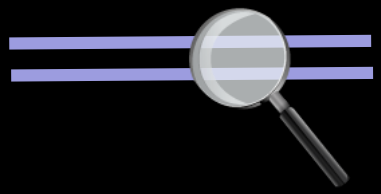
SIDD-signature



Specific non-B DNA structures are accompanied by specific tag distribution

Z-DNA-signature





Summary –

- Enrichment of ssDNA-seq in predicted non-B-DNA structures
- Consistency of tag distribution with theoretical predictions

suggests (abundant) occurrence of non-B DNA structures in *vivo*

Questions

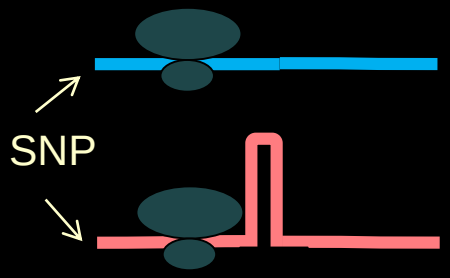
- What is their role, mechanism, interaction with other regulatory elements and Pol II activity

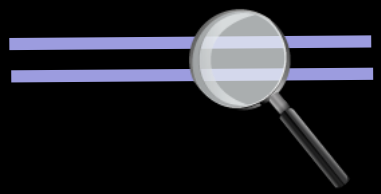
Conformational changes in mRNA structure induced by SNPs

Special acknowledgments: *Raheleh Salari*

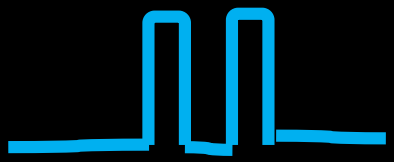
Motivation:

- 5'UTR mRNA structure has regulatory role
- Impact of mRNA structural changes in coding region
  - possible impact on splicing
  - changes in translation dynamics leading to altered folding kinetics of nascent protein chain, and potentially protein misfolding

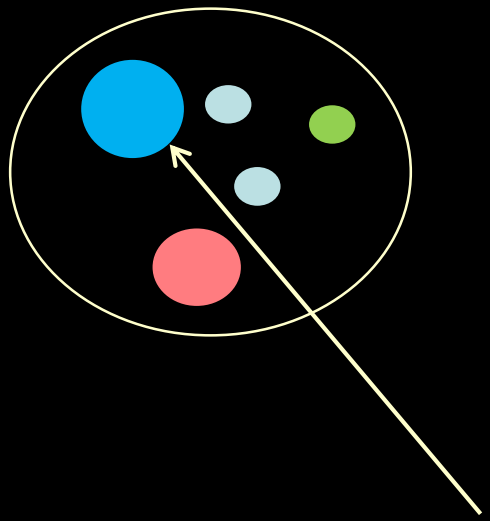




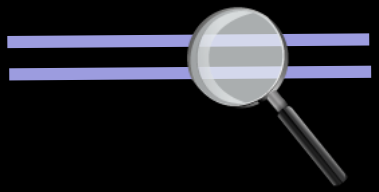
Computational estimation of the impact of a SNP on mRNA structure



p – assembly structures assumed by wild type:

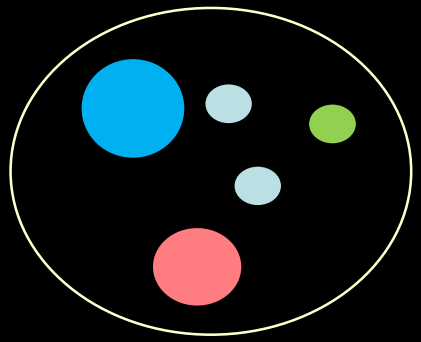


Minimum free energy structure
(area of a circle ~ probability of the structure)

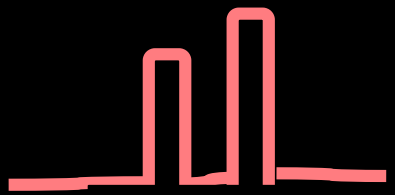


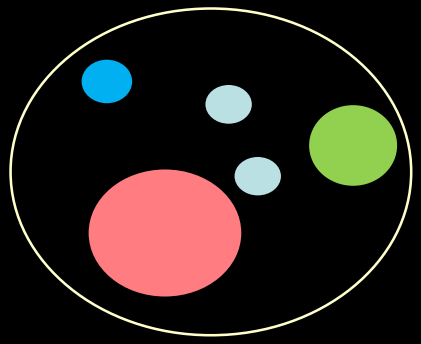
Computational estimation of the impact of a SNP on mRNA structure

assembly structures assumed by wild type



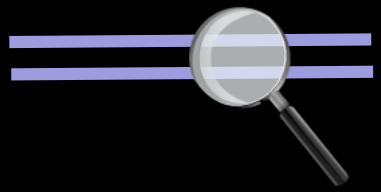
assembly of structures assumed by mutant





Need to compare assemblies of structures taking into account the probability of each structure

!



Measures for comparing of assemblies of structures

p – probability distribution for the assembly of structures assumed by wild type

q – probability distribution of the assembly of structures assumed by the mutant

(a) Relative entropy

$$H(p\|q) = \sum p_s \log(p_s / q_s)$$

(b) Boltzmann distribution distance

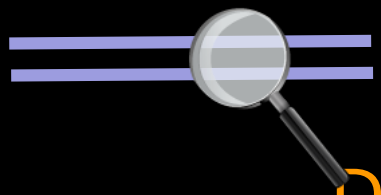
$$d_B(p,q) = \frac{1}{Z_a + Z_b} \sum | e^{-E_{sa}/RT} - e^{-E_{sb}/RT} |$$

(c) Base pair probability distance

$$d_{BP}(p,q) = \frac{1}{n} \sum_{i<j} p_{ij} + q_{ij} - 2 p_{ij} q_{ij}$$

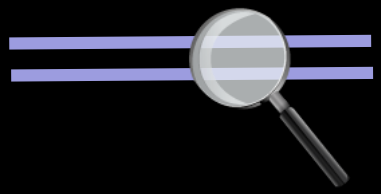
Algorithm using the dynamic programming approach

Salari et al., RECOMB 2011 best poster award
Manuscript in preparation



Disease associated SNPs with significant changes in 5'UTR of mRNA structure

HGMD_ID	gene	type	SNP	Relative Entropy	Boltzmann	BP distance	Disease association
CR063412	PSMA6	5UTR	C102G	24.172115	0.999975	0.340256	**Myocardial infarction**
CR004298	PROC	5UTR	A31C	16.017752	0.998877	0.48465	**Vitamin C deficiency**
CR021383	FMR1	5UTR	G107C	15.988396	0.998711	0.380948	**Fragile X mental retardation syndrome**
CR014434	BDNF	5UTR	C301T	13.781923	0.999308	0.294717	**Alzheimer disease**
CR073552	NIPBL	5UTR	C406T	12.521728	0.990199	0.290129	**Cornelia de Lange syndrome**
CR011064	FTL	5UTR	C14G	11.890619	0.999761	0.382966	**Hyperferritinaemia-cataract syndrome**
CR040301	MX1	5UTR	G88T	11.583429	0.99957	0.251678	**Panencephalitis, sclerosing**
CR900265	HBB	3UTR	A111G	11.047469	0.997338	0.252509	**Thalassaemia beta**
CR025352	ABCA1	5UTR	C35G	10.43122	0.996079	0.452606	**Increased triglyceride level**
CR025435	EGF	5UTR	A71G	10.087699	0.998617	0.22925	**Malignant melanoma**
HR030029	FTL	5UTR	C29G	9.462504	0.979854	0.242743	**Hyperferritinaemia-cataract syndrome**
CR061334	FTL	5UTR	T22G	8.622743	0.996843	0.350765	**Hyperferritinaemia-cataract syndrome**
CR066669	PLTP	5UTR	C47G	8.426413	0.996272	0.295778	**Incr. HDL-C levels**



Can structural changes in mRNA explain disease associated silent SNPs ?

Disease associated “silent SNPs” change mRNA structure:

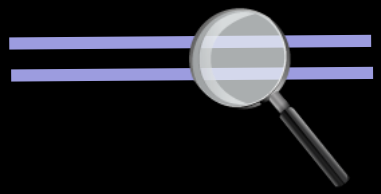
Collaboration with *Andy Fung* (*Michael Gottesman’s* group, CCR)

Disease causing silent mutations in MDR1 gene

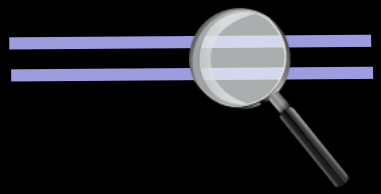
A Silent Polymorphism in the *MDR1* Gene Changes Substrate Specificity –
C. Kimchi-Sarfaty et al. Science 2006

Collaboration with *Chava Kimichi-Sarfaty* (FDA)

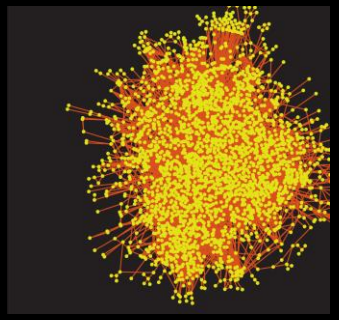
Disease associated silent mutation in F9 - coagulation factor IX gene



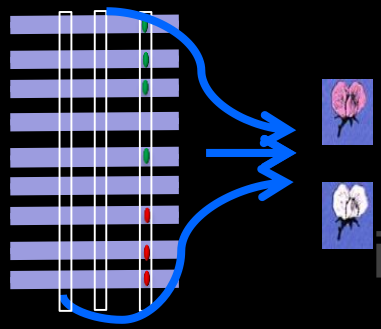
Summary - Efficient computational method to measure effect of SNPs on mRNA allows us to identify disease associated SNPs that have impact mRNA structure



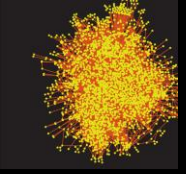
Impact of variations in DNA/RNA sequence structure on gene function



How genetic variations propagate through the molecular system?



Combinatorial regulation and epsitatic interactions



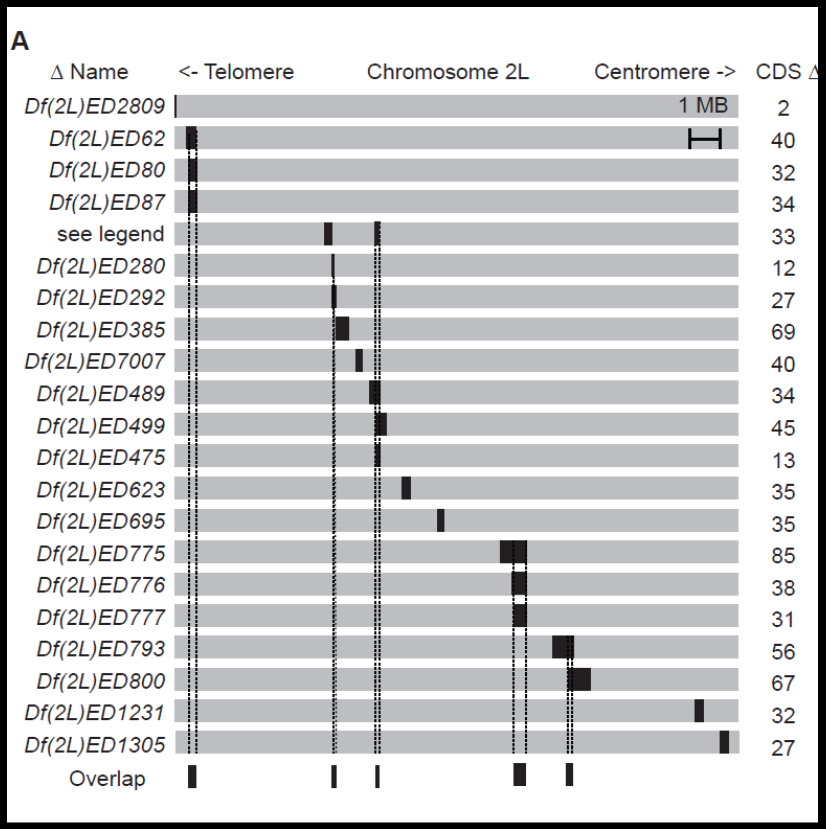
Copy number variations (CNV) (gene dosage)

- implicated in large number of human diseases (cancer, Crohn's disease, autism)
- 28,025 structural variants identified in 1000 genome study (2,000 changes affecting full genes or exons)
- Frequent type of somatic mutations in cancer

Impact of gene dosage on gene expression

(Drosophila melanogaster)

collaboration with *Brian Oliver* group (NIDDK)
Special acknowledgments:
John Malone, *DongYeon Cho*

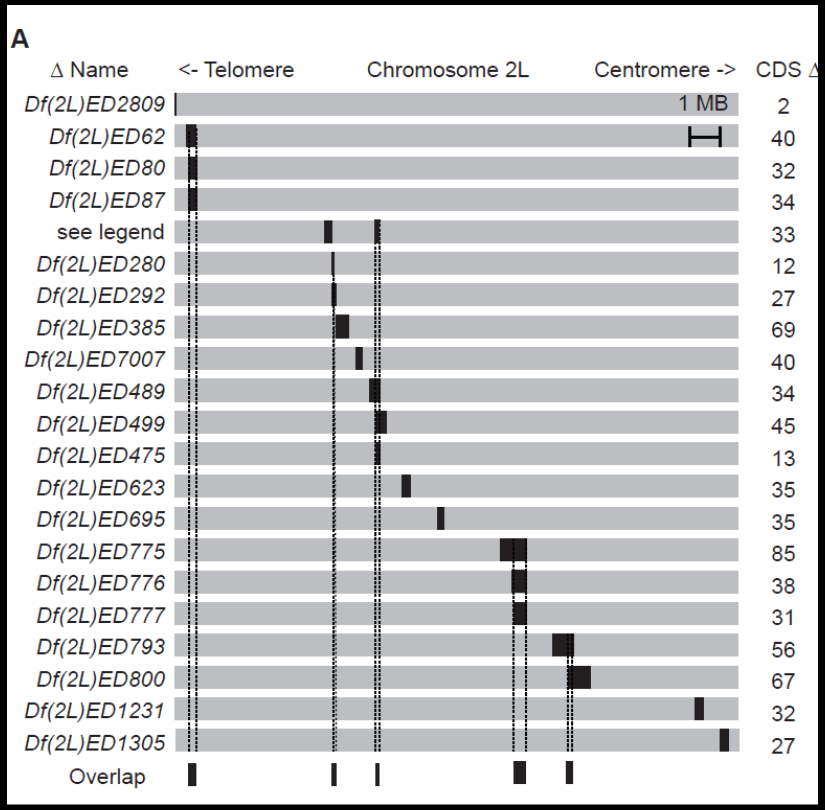


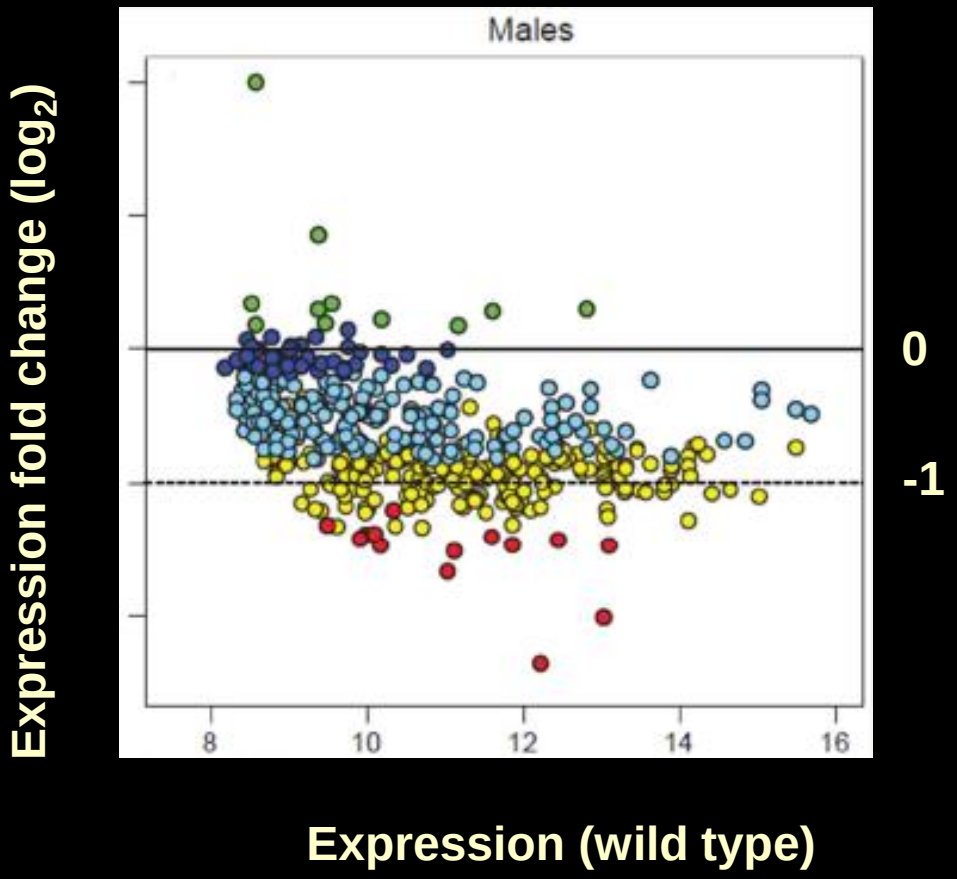
DrosDel lines

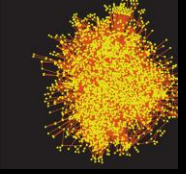
Impact of CNV on gene expression

collaboration with *Brian Oliver* group (NIDDK)

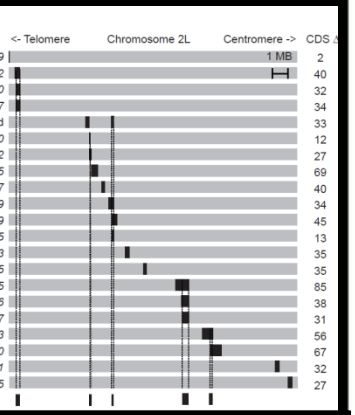
Special acknowledgments:
John Malone, *DongYeon Cho*

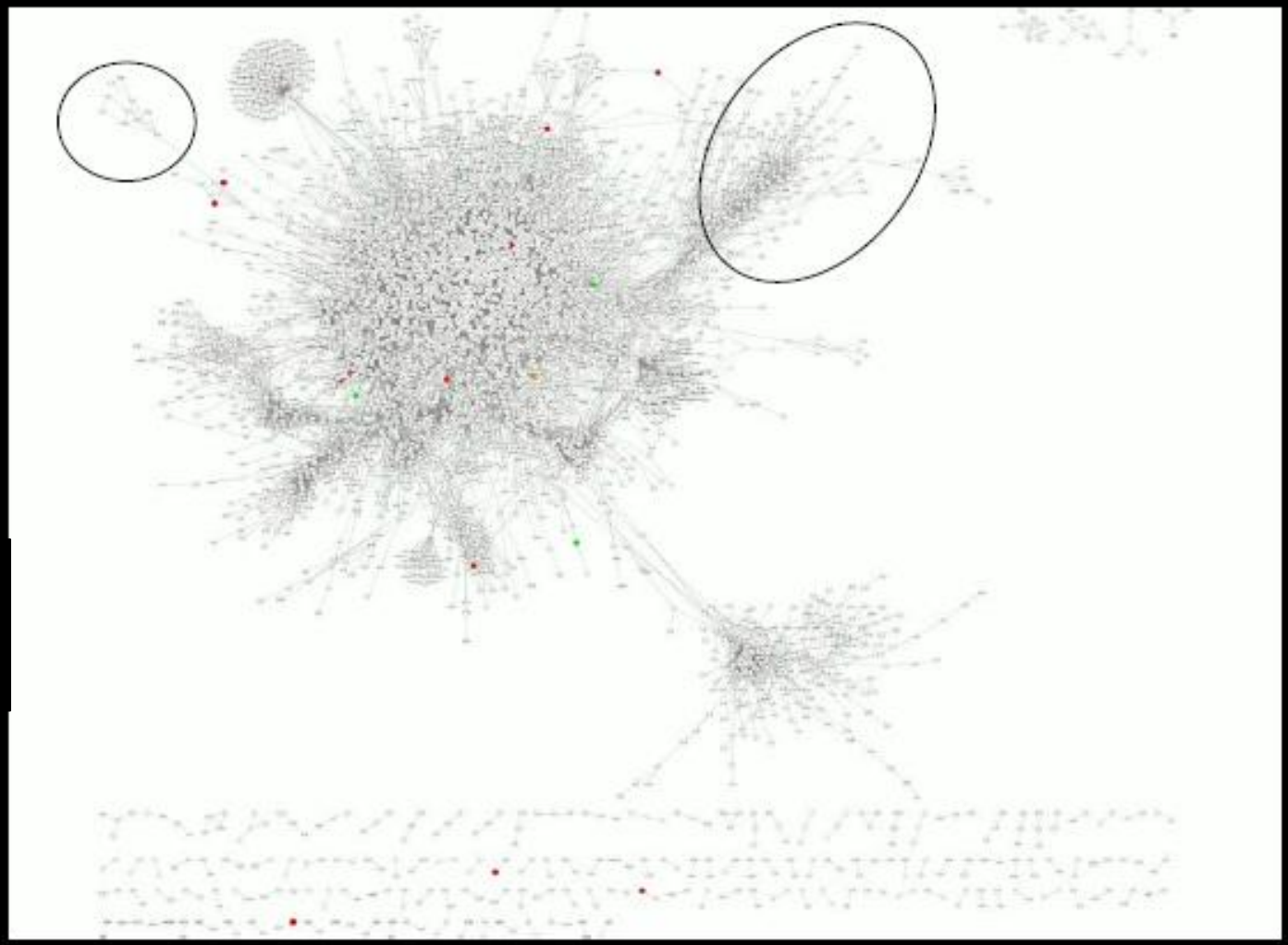






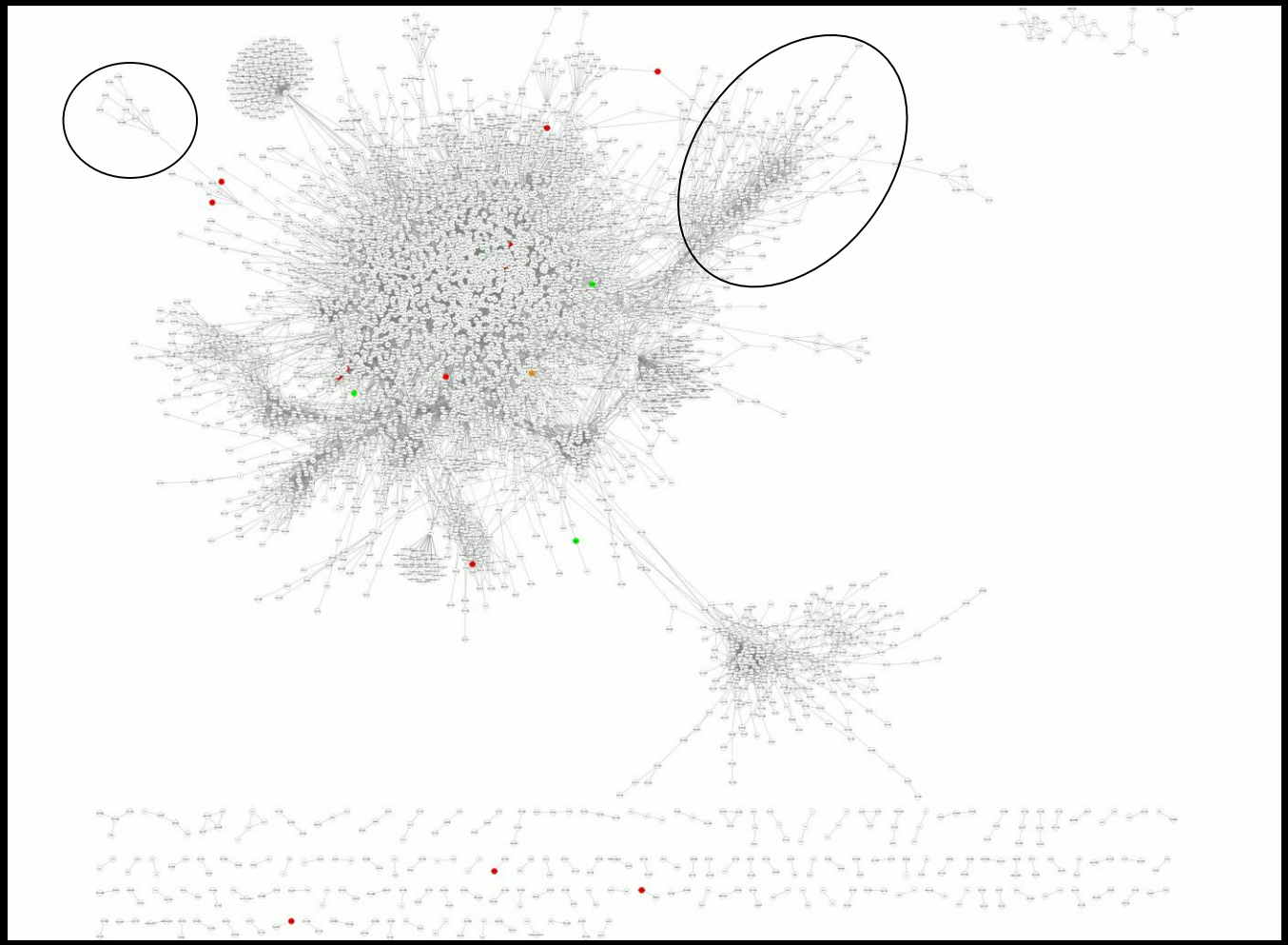
Impact on genes outside the deletions





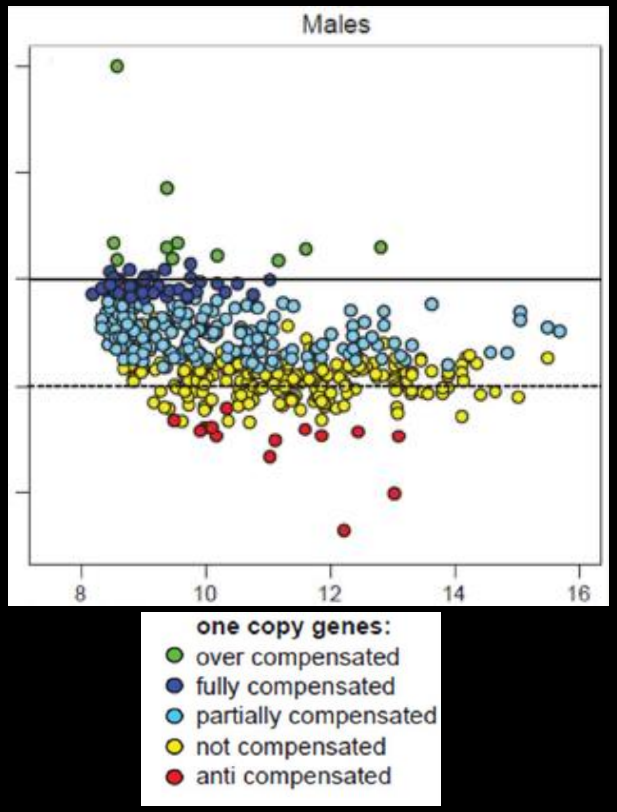
Fly functional network from Costello et.al GB 2009

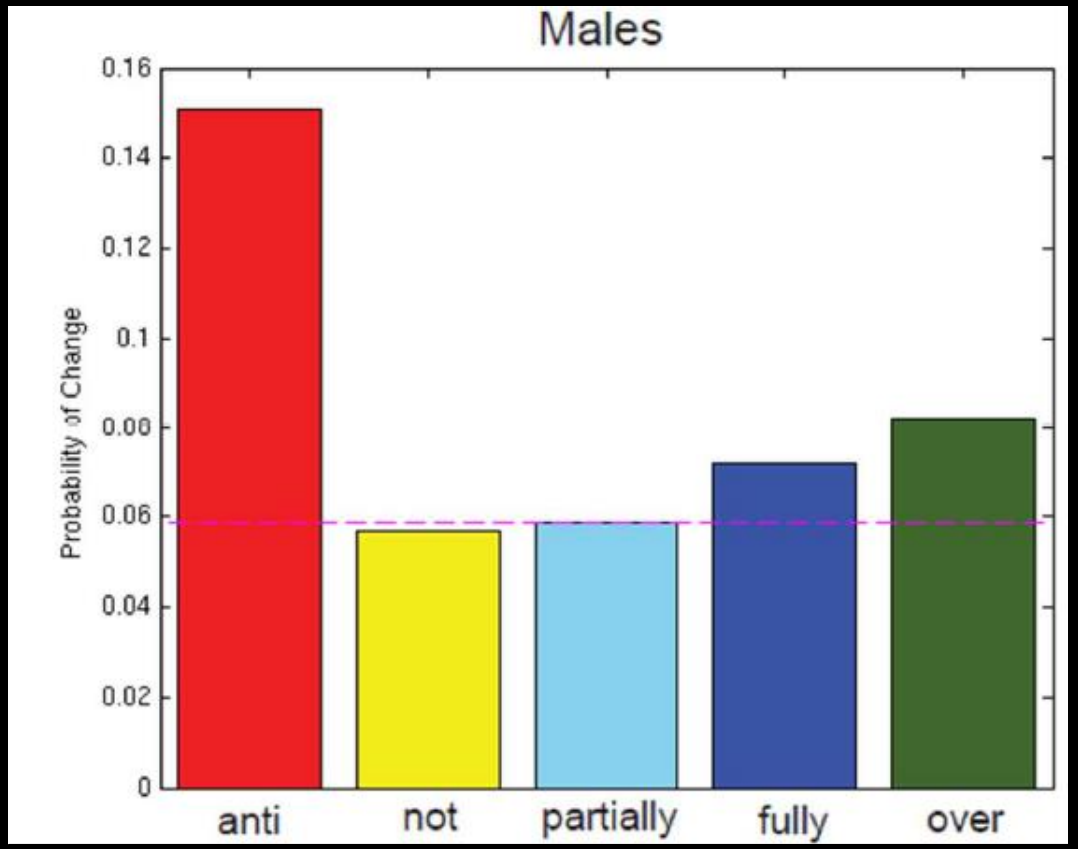
Impact on genes outside the deletions

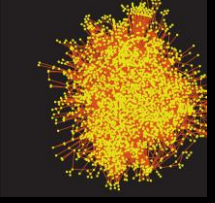


Fly functional network from Costello et.al GB 2009

Genes which show drastic changes due to reduced gene dosage are much more likely to see expression changes for 1st order network neighbors



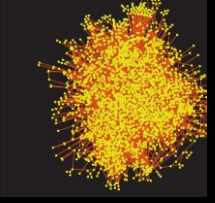




Summary:

Copy number variations impact expression of genes outside deletion in network-dependent way

Indication of feedback loops acting on deleted genes

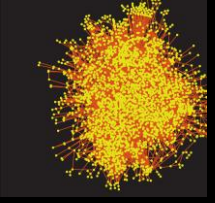


Propagation of the effects of CNV in Glioma

Special acknowledgments:
Yoo-ah Kim

Motivation:

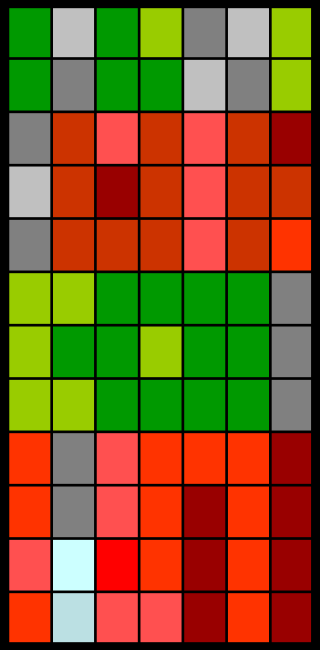
- A method for system level analysis of propagation of such perturbation in the network
- Identification pathways dys-regulated in disease
- Capturing master regulators (network hubs) involved in disease
- Identification “causal” mutations



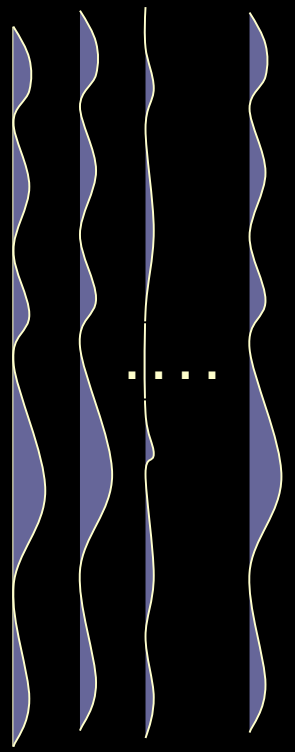
Propagation of the effects of CNV in Glioma

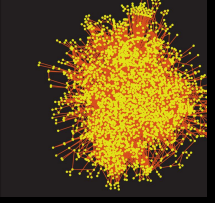
Cancer Cases
Gene expression data

Gene 1
Gene 2
Gene 3
.
.
.
.
.
Gene n



Cancer Cases
CNV data

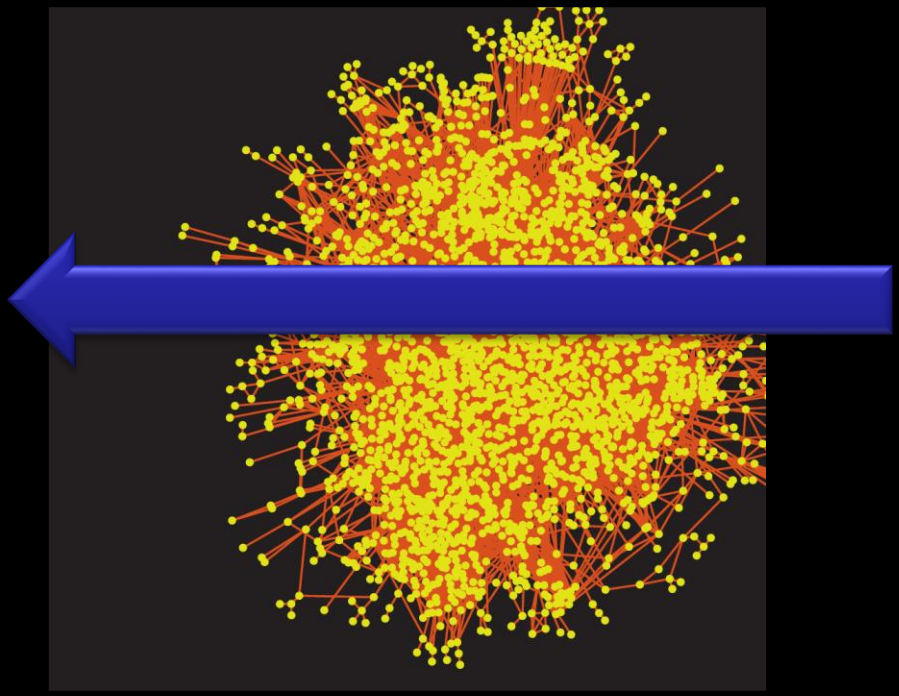




Propagation of the effects of CNV in Glioma

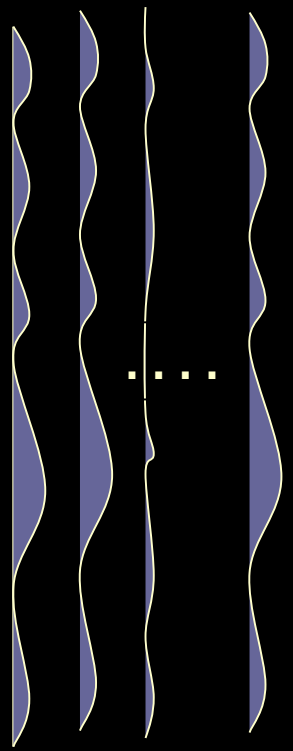
Cancer Cases
Gene expression data

Gene 1
Gene 2
Gene 3
.
.
.
.
.
Gene n



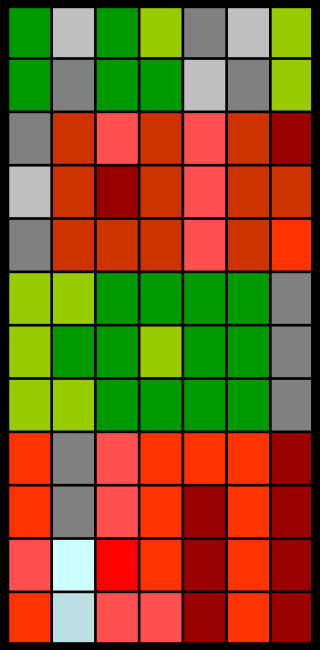
Integrated
Protein-protein, protein-DNA
phosphorylation network

Cancer Cases
CNV data

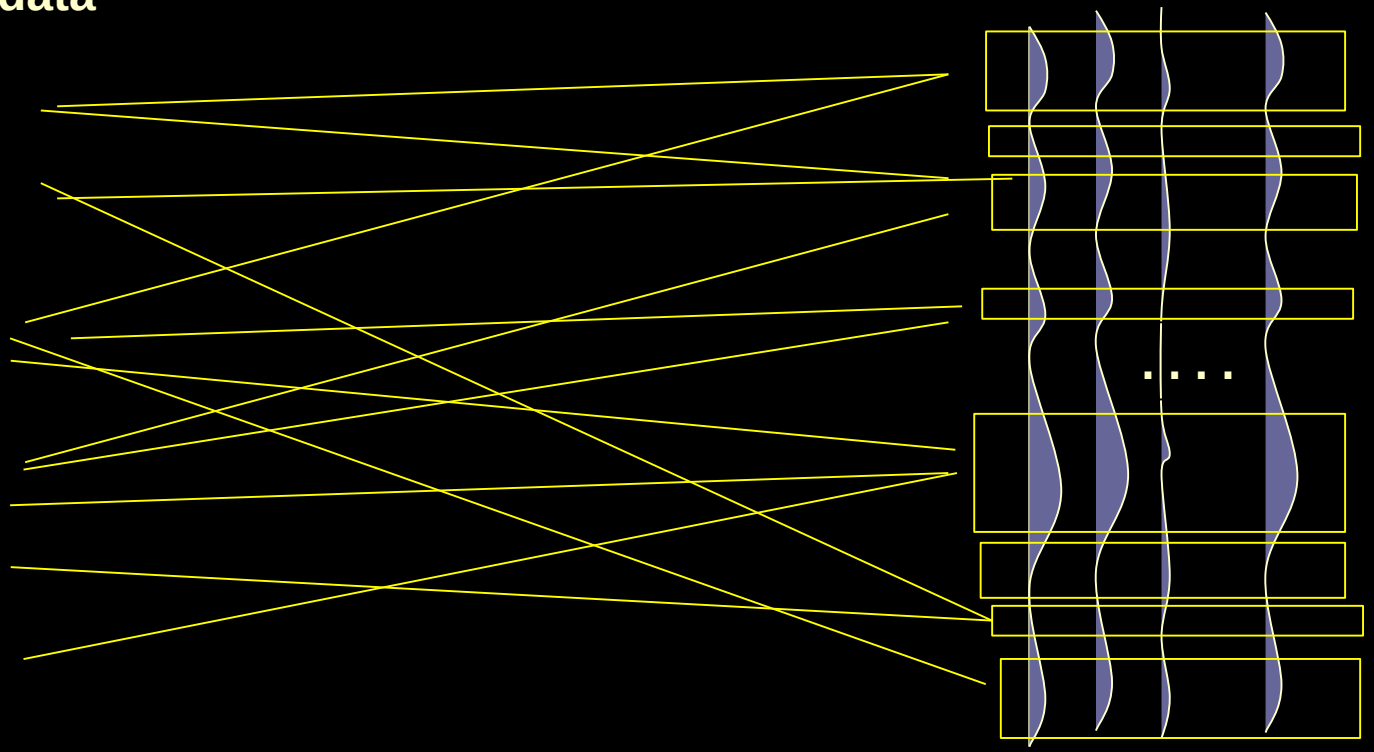


eQTL analysis : Testing associations between copy number variations and gene expression

Cancer Cases
Gene expression data

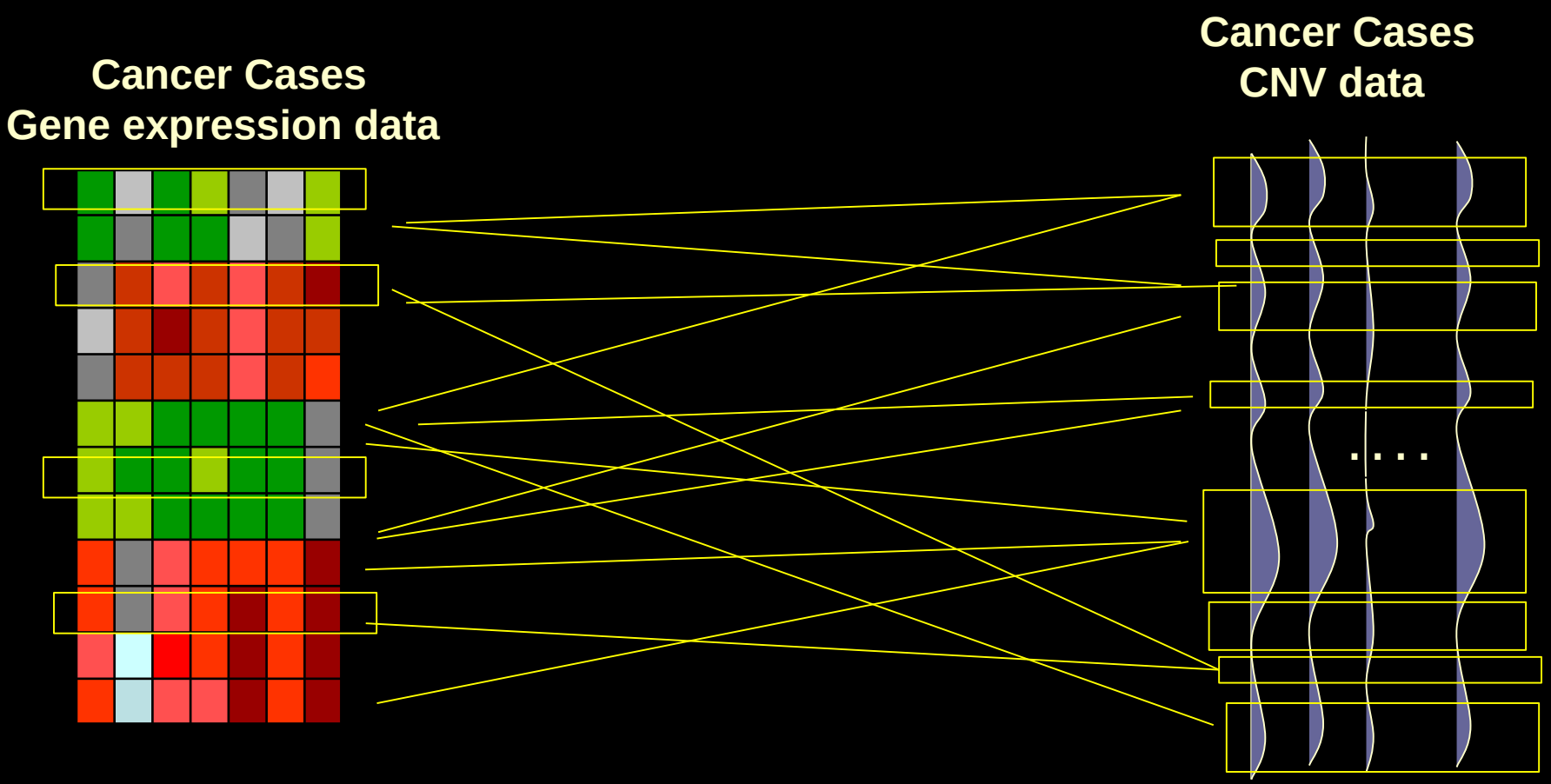


Cancer Cases
CNV data

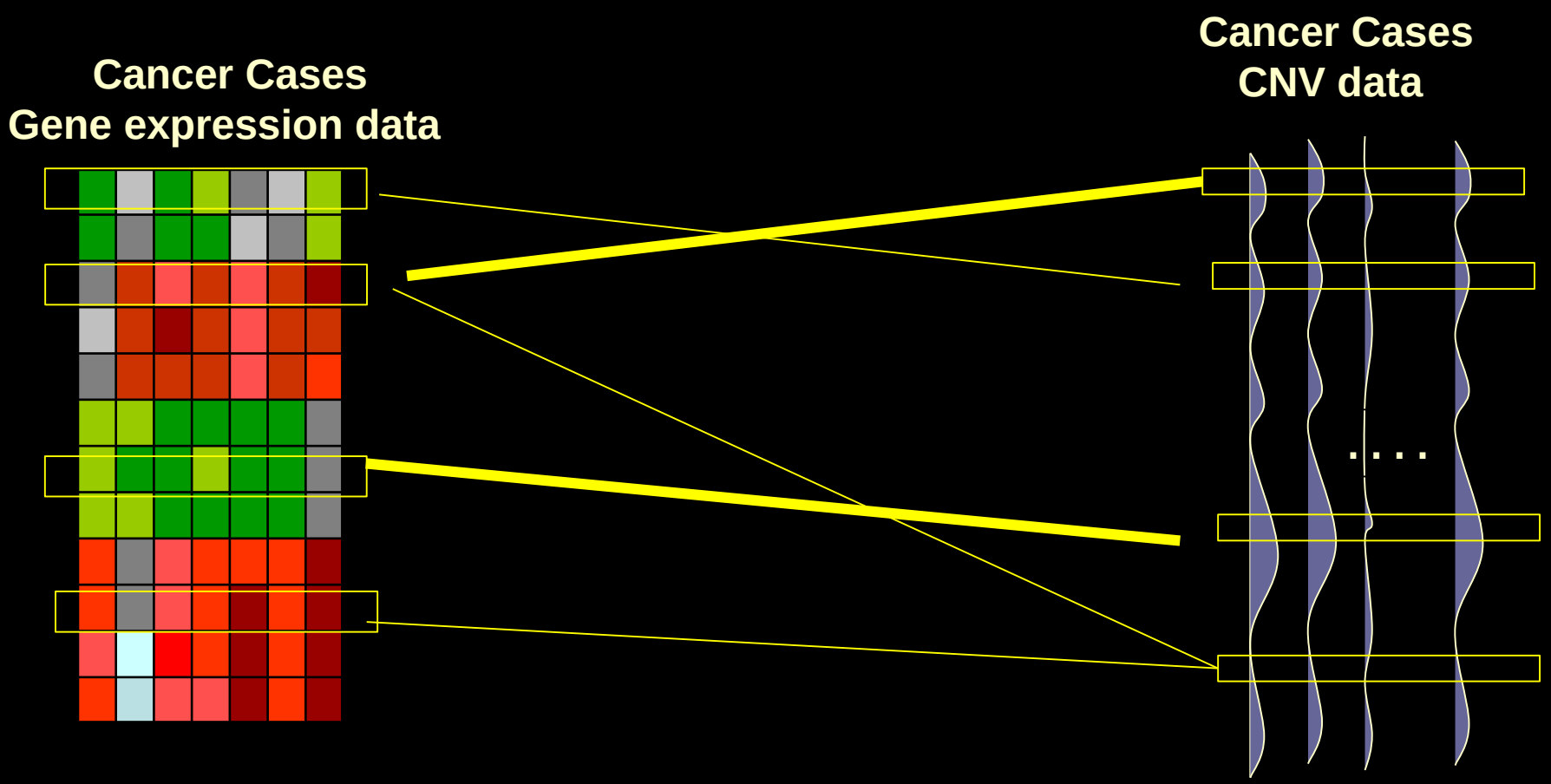


Too many possible pairs to consider them all !!!

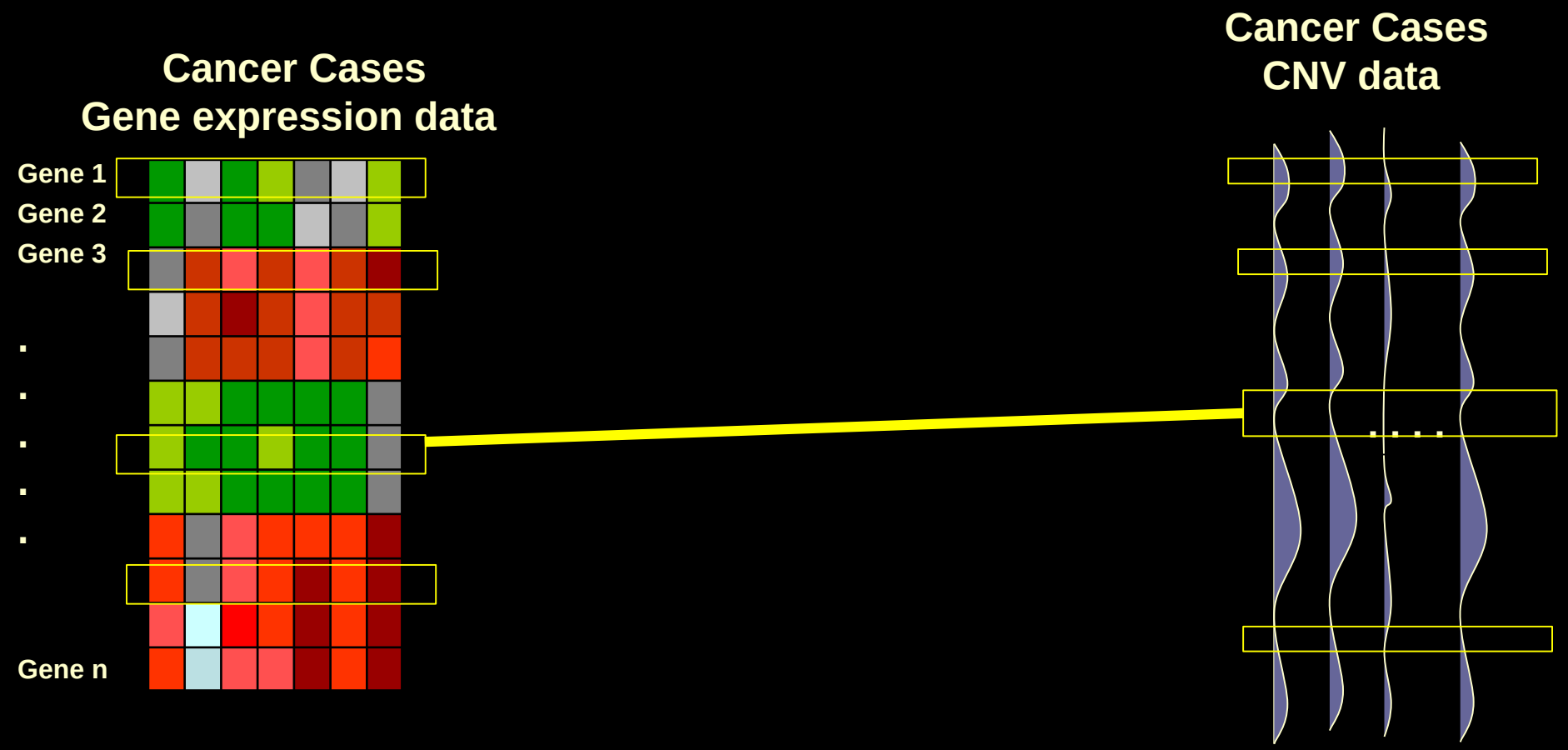
Selecting representative set of target genes



Associations between copy number variations and gene expression of selected target genes

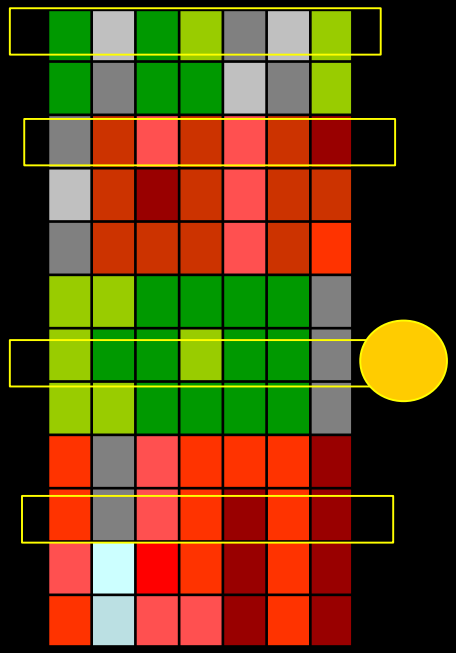


Uncovering pathways of information flow between genetic loci and a target gene

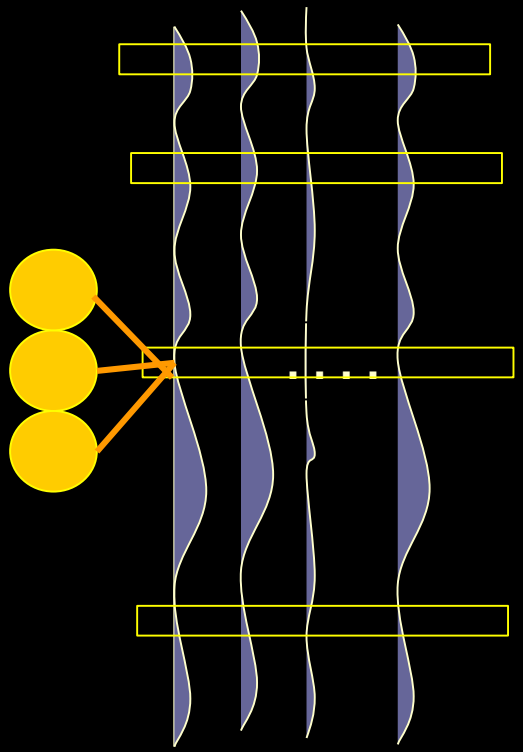


Uncovering pathways of information flow between CNV and target gene

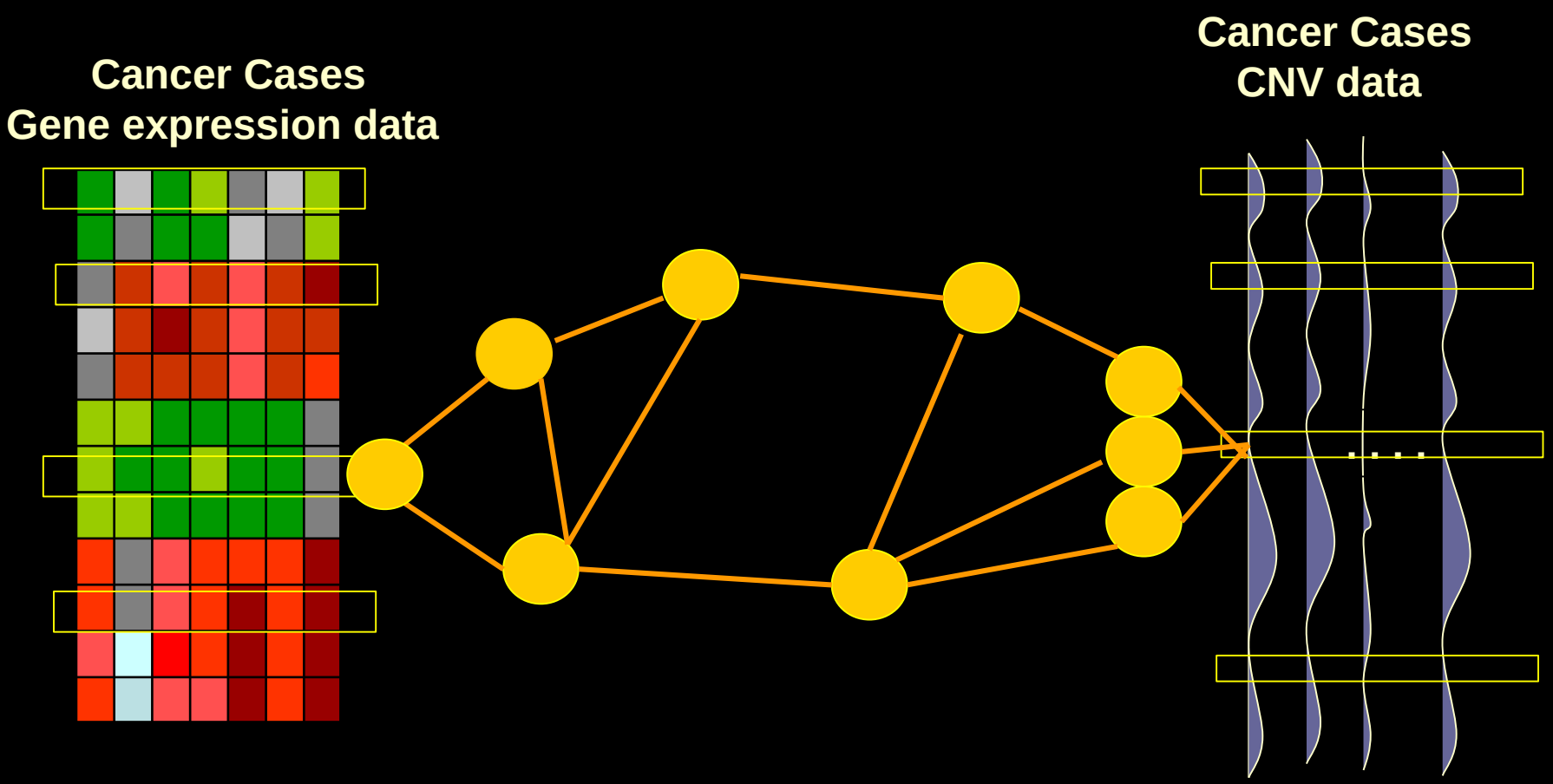
Cancer Cases
Gene expression data



Cancer Cases
CNV data



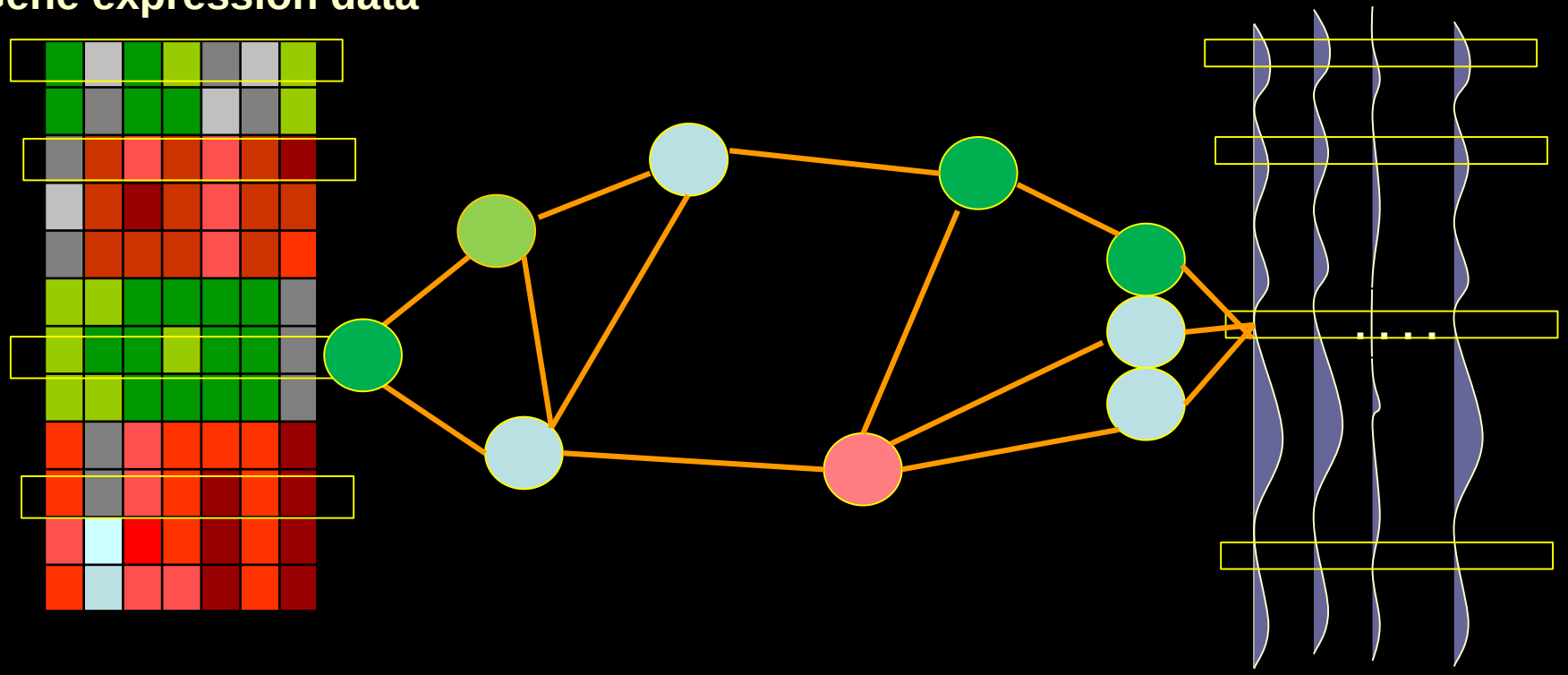
Uncovering pathways of information flow between CNV and target gene



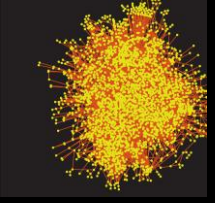
Uncovering pathways of information flow between CNV and target gene

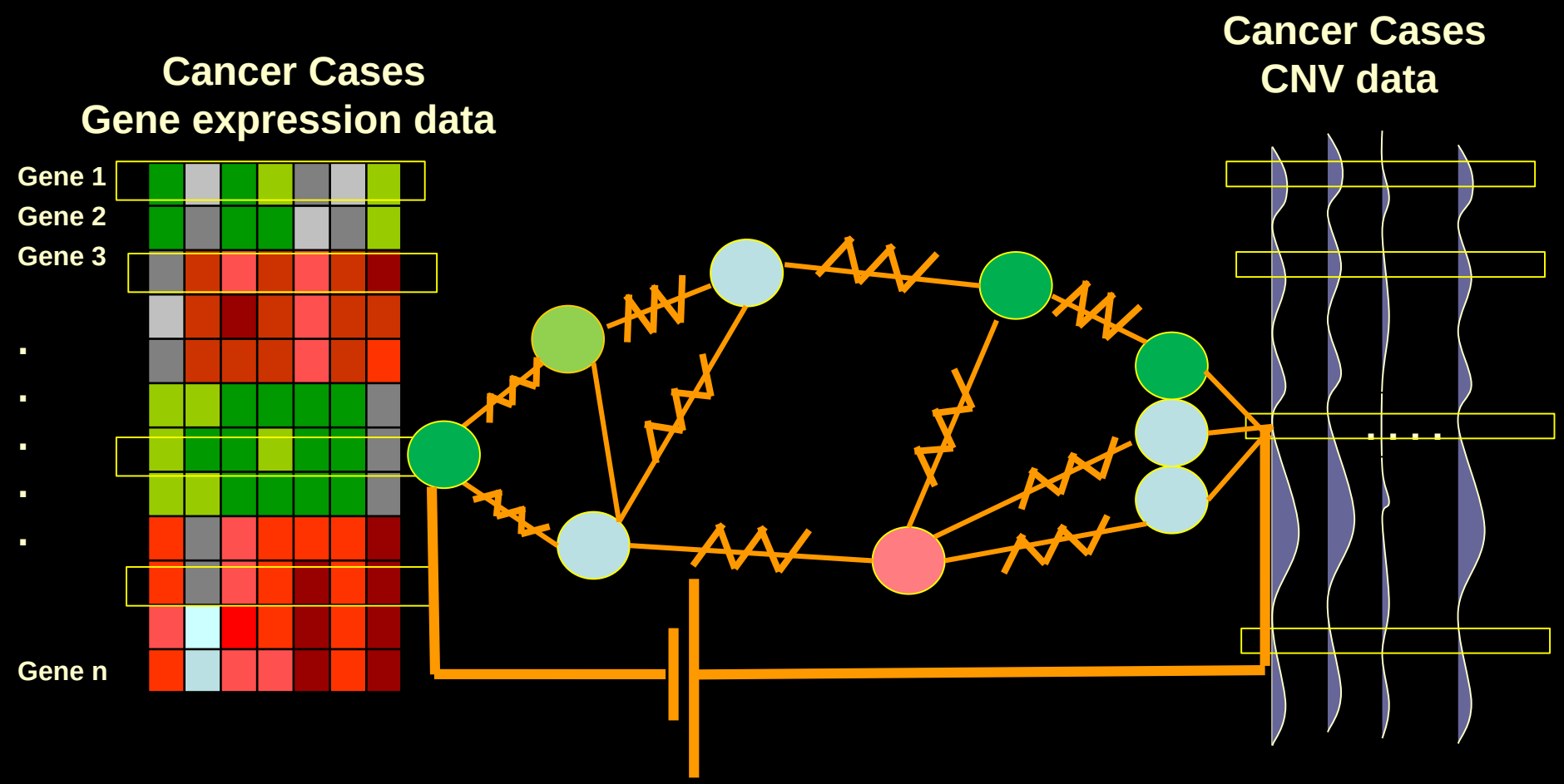
Cancer Cases
Gene expression data

Cancer Cases
CNV data



Uncovering pathways of information flow between CNV and target gene

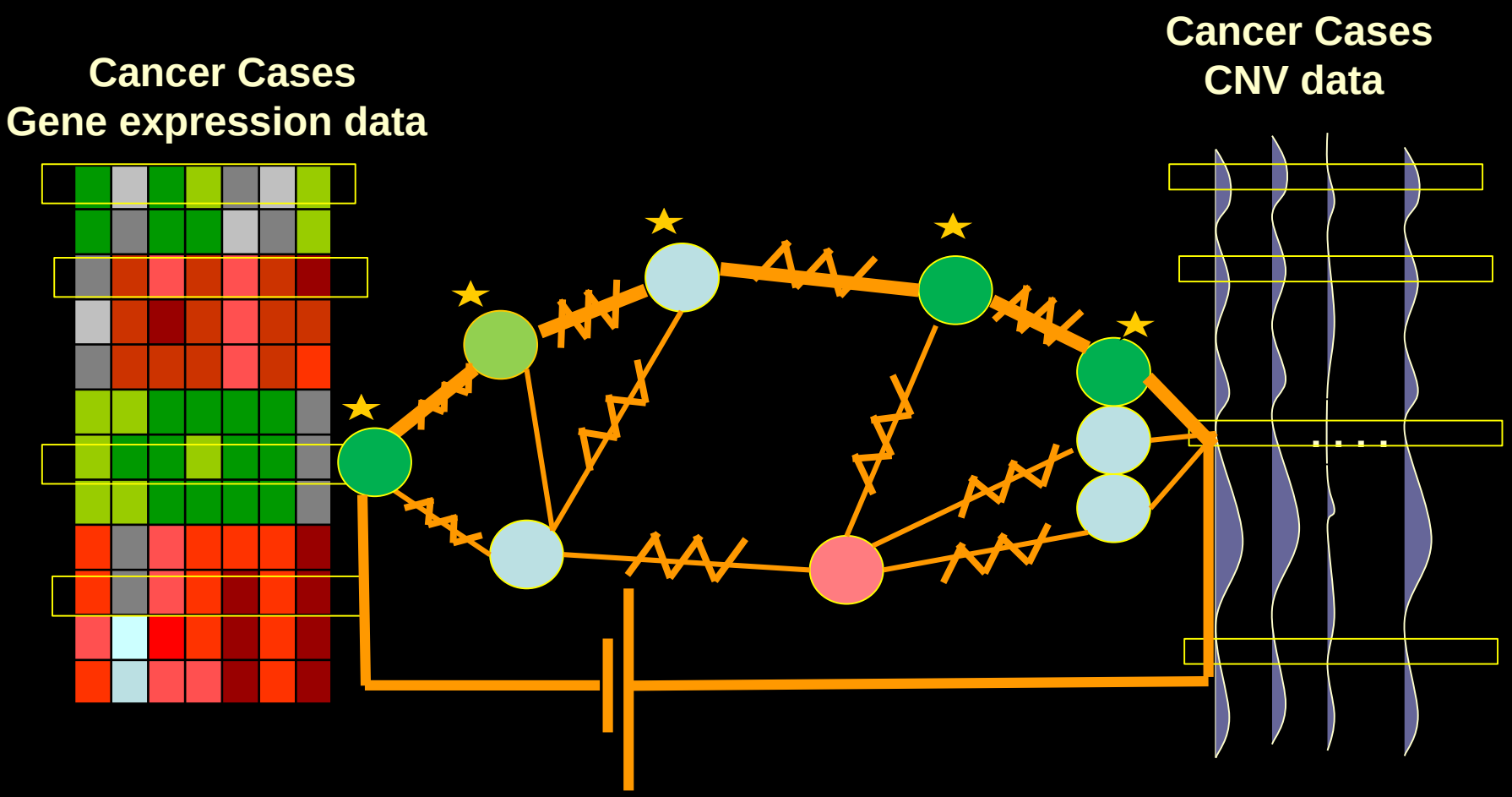




Resistance **- set to favor most likely path -based on gene expression values**

(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

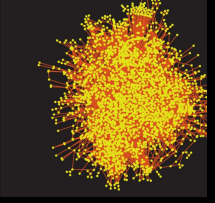
Possible means of information flow between CNV and target gene?

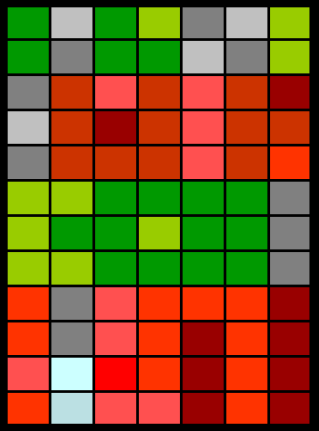


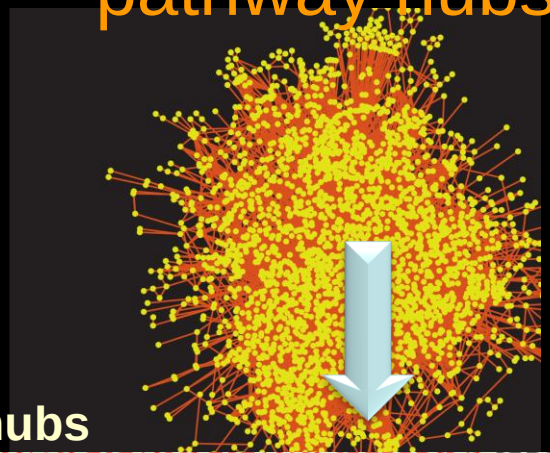
Resistance - set to favor most likely path -based on gene expression values

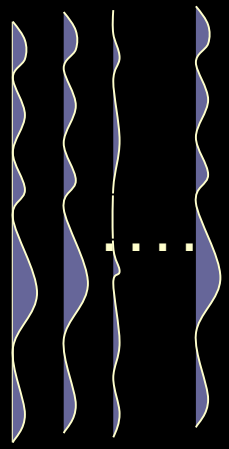
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

Selected target genes, causal genes and pathway hubs









hubs

MYC(110) E2F1(88) E2F4(43) CREBBP(34) GRB2(27) SP3(26) ESR1(25) TFAP2A(25) NFKB1(23) MYB(22) JUN(22) E2F2(22) RELA(21) AR(21) SP1(20) RPS27A(20) MAPK3(19) POU5F1(17) HIF1A(16) PPARA(15) CDC42(15) UBA52(13) CDK7(13) YBX1(13) YWHAZ(12) CEBPB(12) POU2F1(12) UBE2I(11) SMAD3(11) TAL1(11)

Target genes

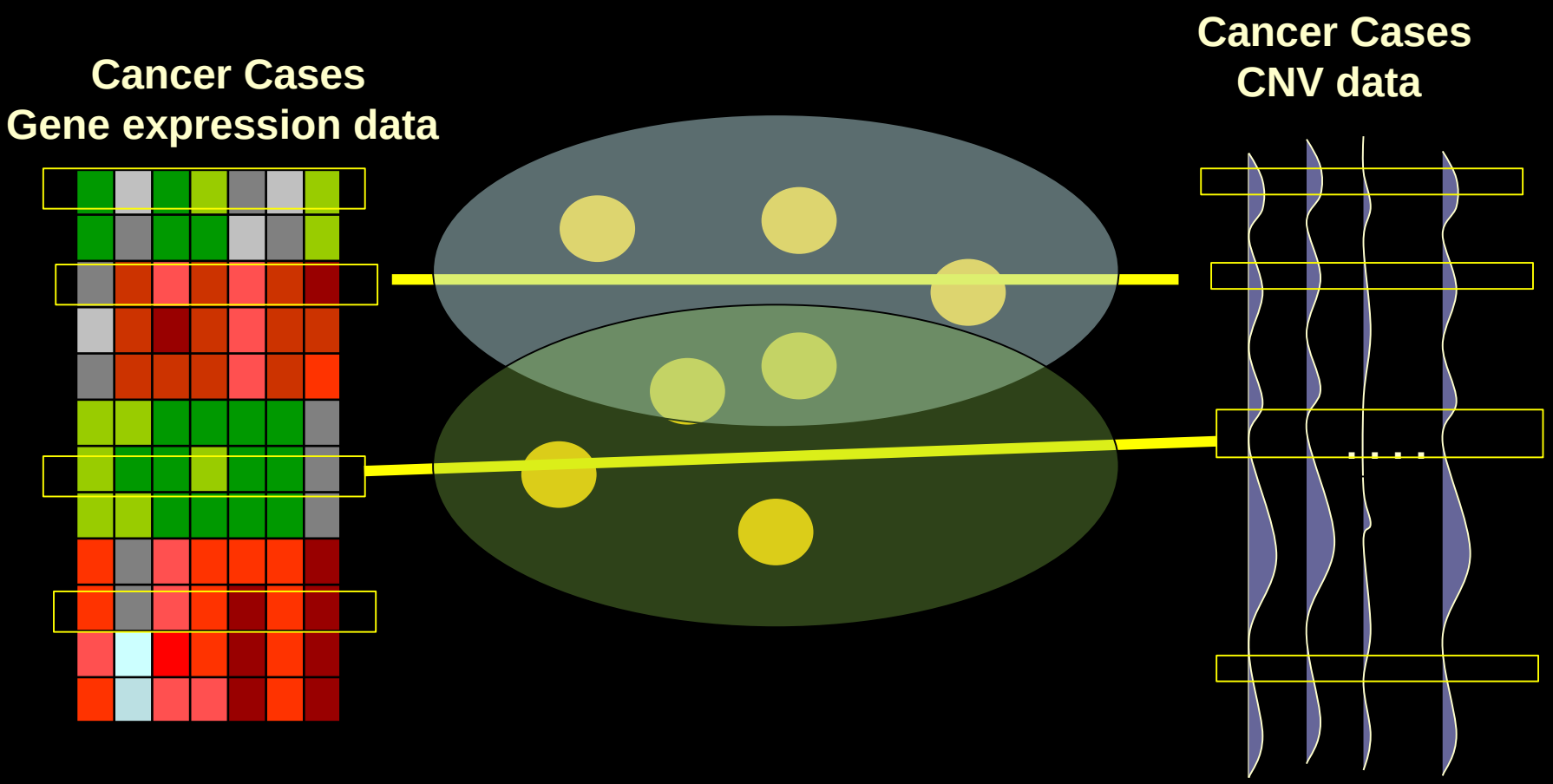
ABCA1 ACCN1 ADRBK2 ANK3 ANKRD26 ANXA2 APLP1 ARHGDIG ATP6V1G2 BCAT1 BTG1 CD163 CDK2 CDK4 CDKN2D CFI CHEK1 CNIH4 CSDA CTSK CXCR4 DOCK9 EGFR ETS1 F2R FLNA FN1 GABRA4 GBE1 GBP1 GNS GOT1 HEXB HMGB2 HTR2A IGFBP2 IGFBP3 LAMC1 LPL MAP1A MCM3 MDK MMP2 MSN NELL1 PCNA PEG3 PLAU PLSCR1 PPP2R2C PRRX1 PTX3 RAB3A RBBP8 SHC1 SHMT2 SMC4 SNAP91 SNRPG SSTR1 STMN1 STXBP1 SYNGR1 TCF3 TGFB2 TNFRSF10B TP53 TP53I3 TPRKB TRIM22 TSPAN6 UBE2C VAMP2 WEE1

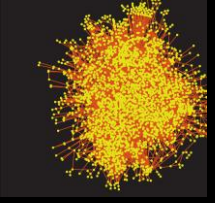
causal genes

ABCA1 ACP1 ADCY8 AGA AHR AKAP6 AKAP9 AKT1 ANXA11 ANXA2 APP ARHGAP11A ARHGAP29 ATR BUB3 CAD CAMK2G CCNC CDC2 CDC5L CDKN2A CEBPA CEP70 CFH CHUK COBL CRMP1 CSF2 CSNK2A1 CUL1 DARC DDX56 DIAPH3 DLC1 EFNA5 EGFR EIF2B1 EIF3A EIF3B EIF3F ELMO1 EPB41 ERBB4 ERCC6 FAS FER FHL2 GBAS GBE1 GSTK1 HEATR1 HSDL2 IFNA4 ILK ITGB3BP KITLG LMO7 MAP2K4 MCM7 MED10 MON2 MRLC2 MS4A1 NDUFA4 NDUFB8 NRXN1 NUP205 NUPL1 ORC5L PARP1 PCDH7 POLR1A POLR2J POLR3A POLR3B POM121 PPIA PRIM1 PRKAB1 PRKCA PSAP PSMA1 PSMA4 PSMA5 PSMB1 PSMC3 PSMC6 PTEN PTK2B PTPRD PTPRJ PTPRK RAI14 RB1 RBMX RPOMS REL RGL1 RHOBTB2 RPL10 RPL10L RPS17 SEC61A2 SF3B4 SFRS2 SFRS3 SGCB SLC25A4 SLC27A2 SNRPB2 SPTA1 STXBP6 SYNGR1 TAF2 TERF2IP THBS1 TOP1 TP53 TRIP13 TSSC1 U2AF2 UBE3A USF2 VAV3 VDAC2 VIM VWF ZNF107

Examples of genes known to be associated with glioma

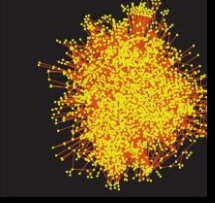
Are there common functional pathways?





Common GO pathways

cell cycle arrest	10
epidermal growth factor receptor signaling pathway	9
negative regulation of cell growth	9
Ras protein signal transduction	9
regulation of sequestering of triglyceride	8
cell proliferation	7
nuclear mRNA splicing, via spliceosome	7
regulation of cholesterol storage	7
nucleotide-excision repair	7
RNA elongation from RNA polymerase II promoter	7
insulin receptor signaling pathway	6
transcription initiation from RNA polymerase II promoter	6
N-terminal peptidyl-lysine acetylation	5
phosphoinositide-mediated signaling	5
positive regulation of lipid storage	4
positive regulation of specific transcription from RNA polymerase II promoter	3
positive regulation of epithelial cell proliferation	3
base-excision repair	2
negative regulation of hydrolase activity	2
gland development	2
positive regulation of MAP kinase activity	2
regulation of nitric-oxide synthase activity	2
estrogen receptor signaling pathway	2
regulation of receptor biosynthetic process	2
response to organic substance	2
JAK-STAT cascade	2
regulation of transforming growth factor-beta2 production	2
G1/S transition of mitotic cell cycle	2
SMAD protein nuclear translocation	2



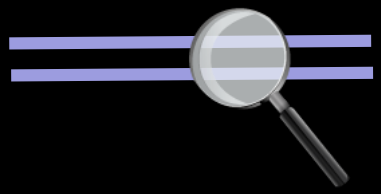
Advantages of current flow approach

- Ability to use expression to guide the flow of information
- Efficiently solvable by solving a set of linear equations (Kirchhoff's laws)
  Caveat: solving a linear system with 20,000 variables thousands of times (permutation test !) required new methodology

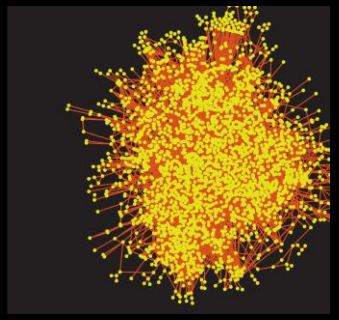
Kim, Przytycki, Wuchty, Przytycka – *Phys. Bio.* – in print

- Extendable to directed edges
- The first approach that addresses the question of relation between CNV and dys-regulated genes/pathways identifying common dys-regulated pathways and causal genes

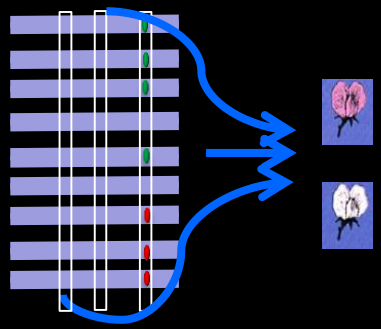
Kim, Wuchty, Przytycka – *PloS Comp Bio 2011*



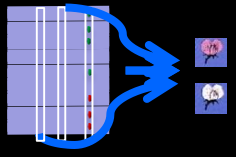
Variation in DNA sequence and structure and their impact on gene expression



How genetic variation propagate through the molecular system?

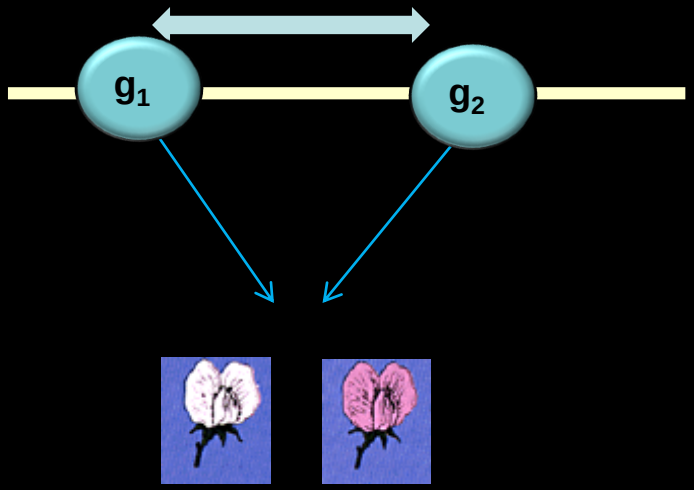


Uncovering epsitatic interactions



Epistatic interactions

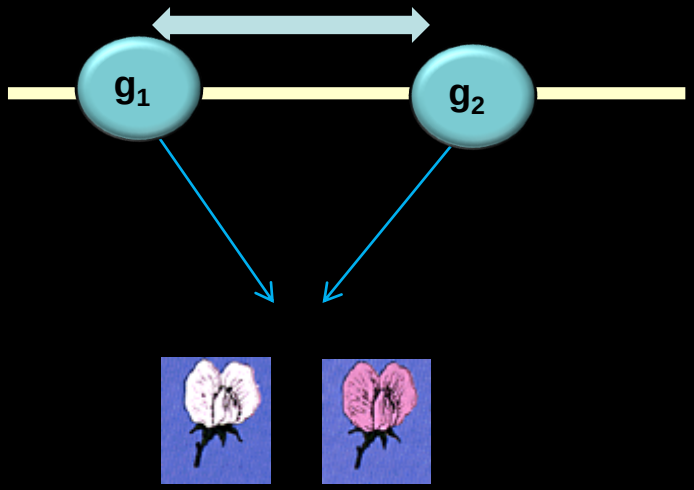
epistatic interaction



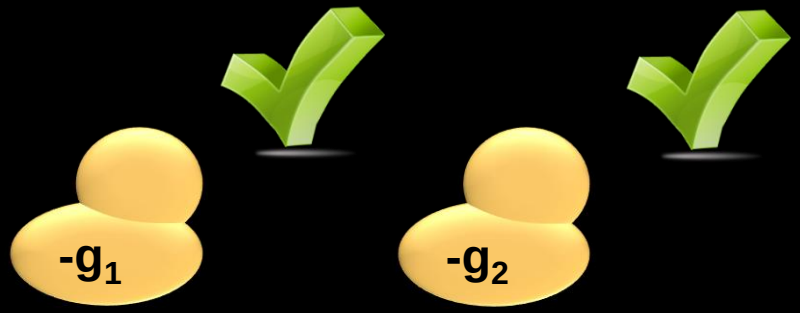
Effect of gene g_1 depends on gene g_2

Epistatic interactions

epistatic interaction



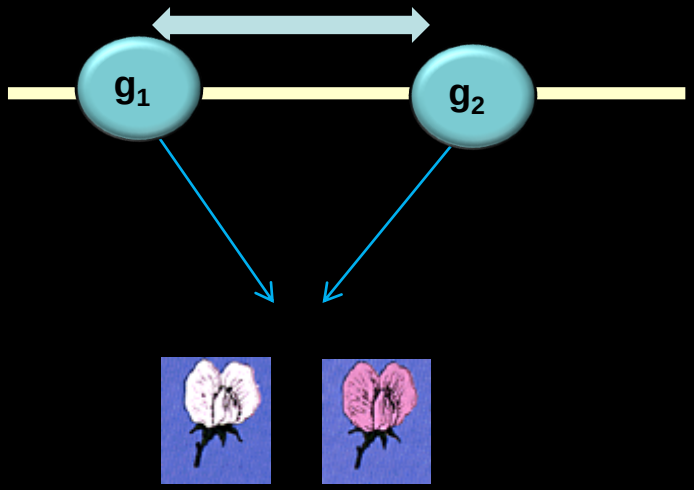
Synthetic lethality



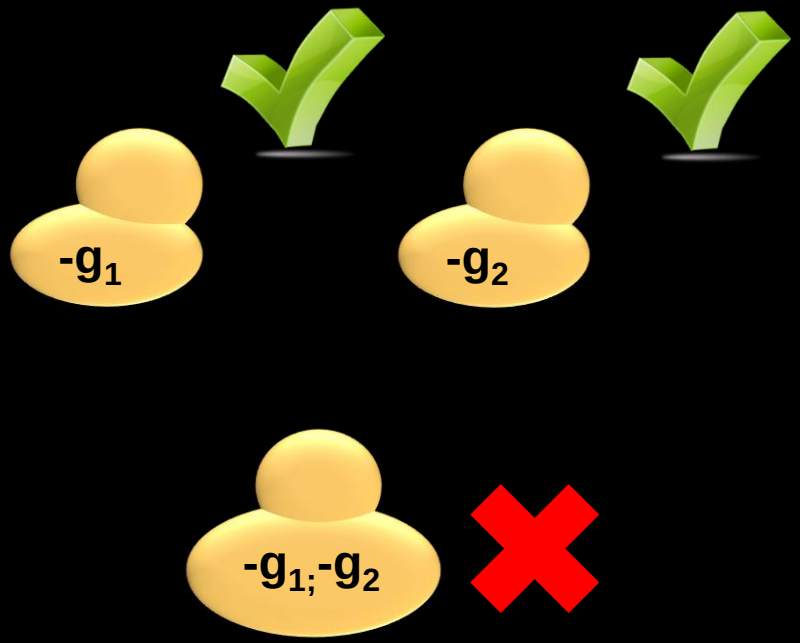
Effect of gene g_1 depends on gene g_2

Epistatic interactions

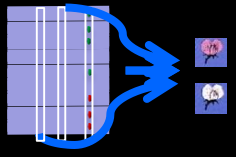
epistatic interaction



Synthetic lethality



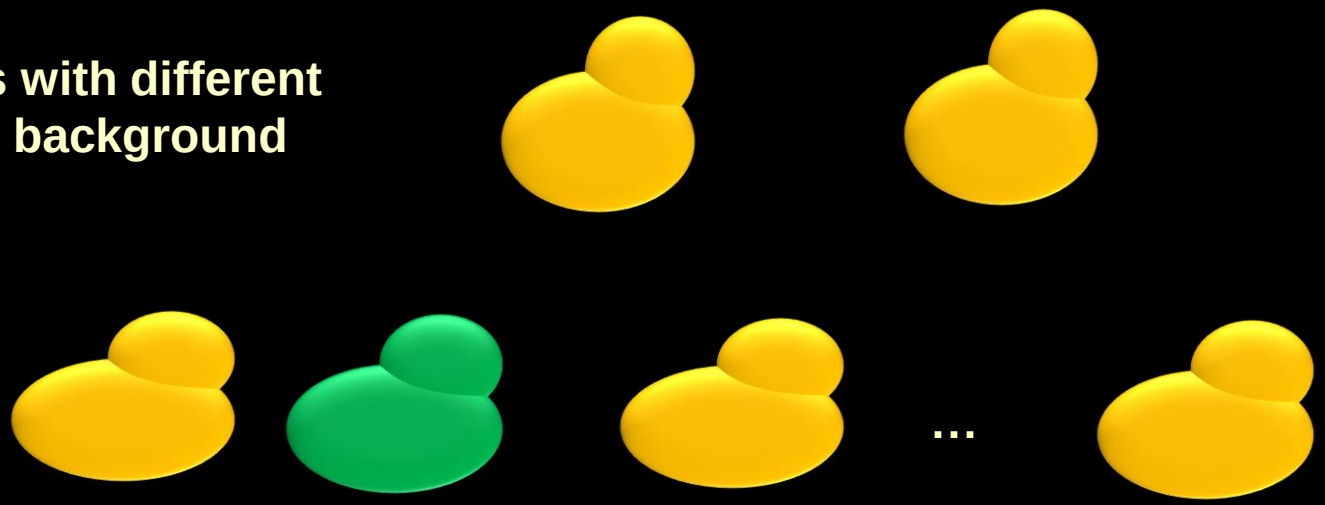
Effect of gene g_1 depends on gene g_2



Uncovering epistatic interactions in the context of genetic crosses

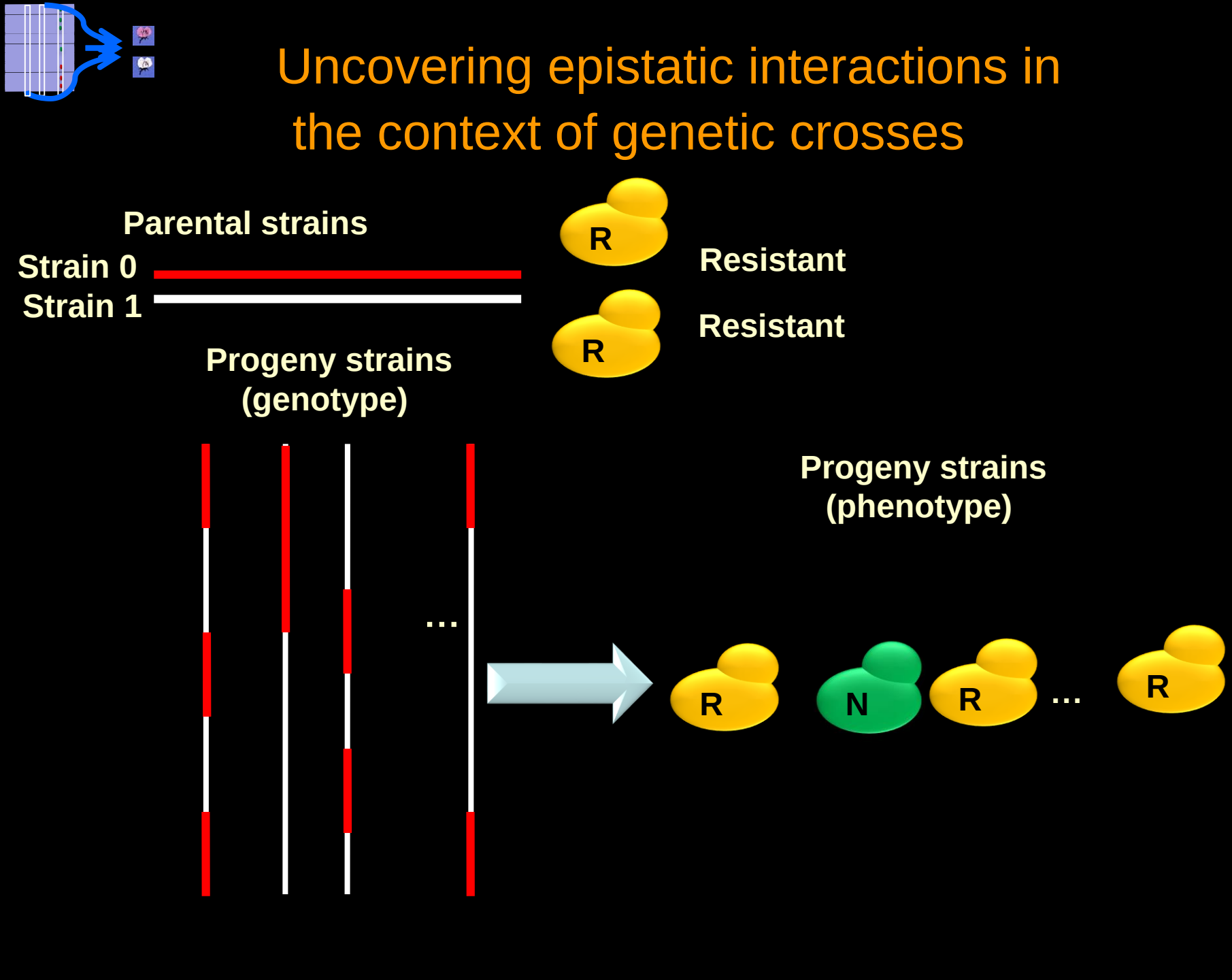
Special acknowledgments:
Yang Huang

Parents with different genetic background

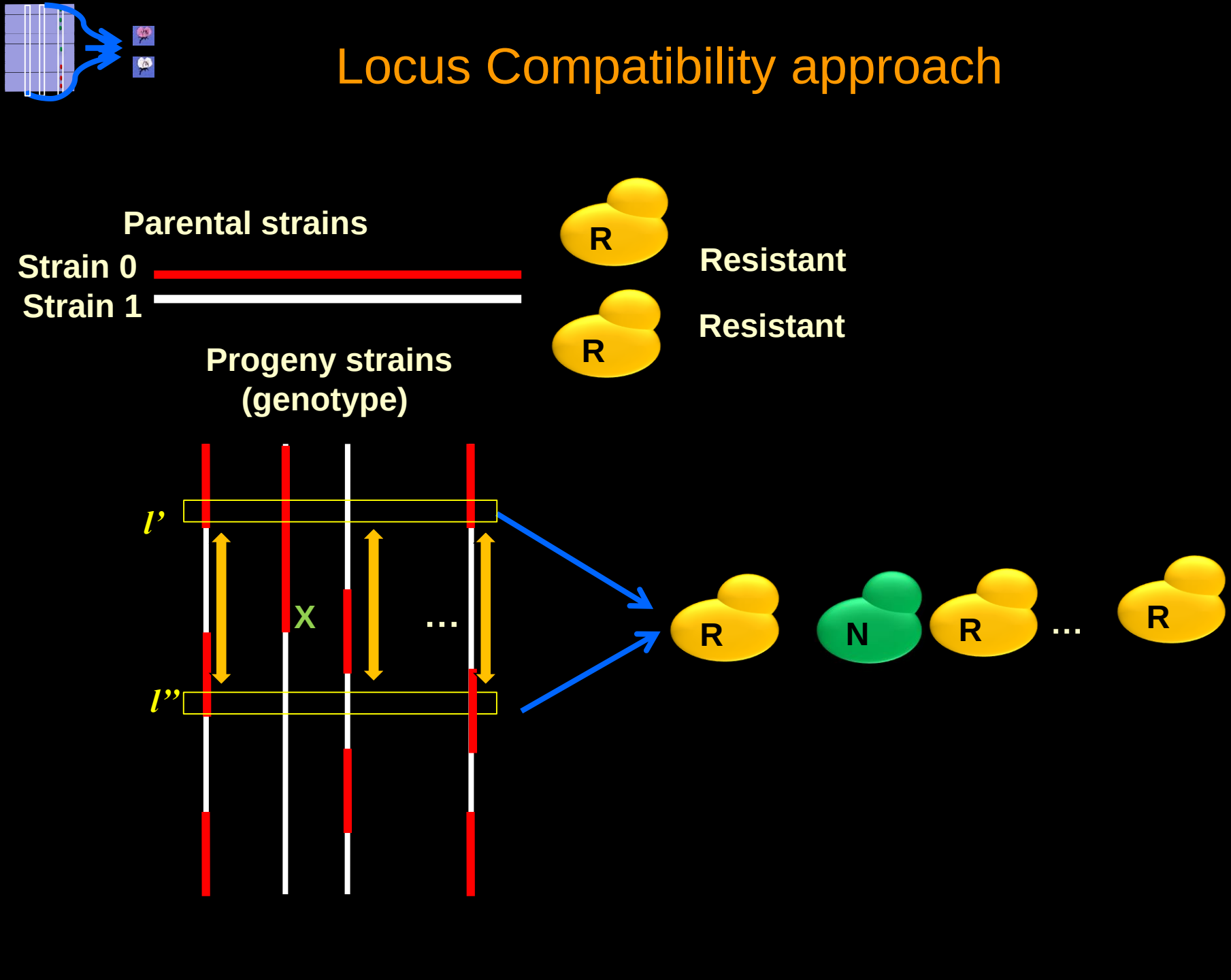


Progenies with different phenotypes

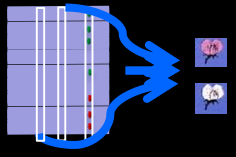
Uncovering epistatic interactions in the context of genetic crosses



Locus Compatibility approach



Locus Compatibility approach



Parental strains

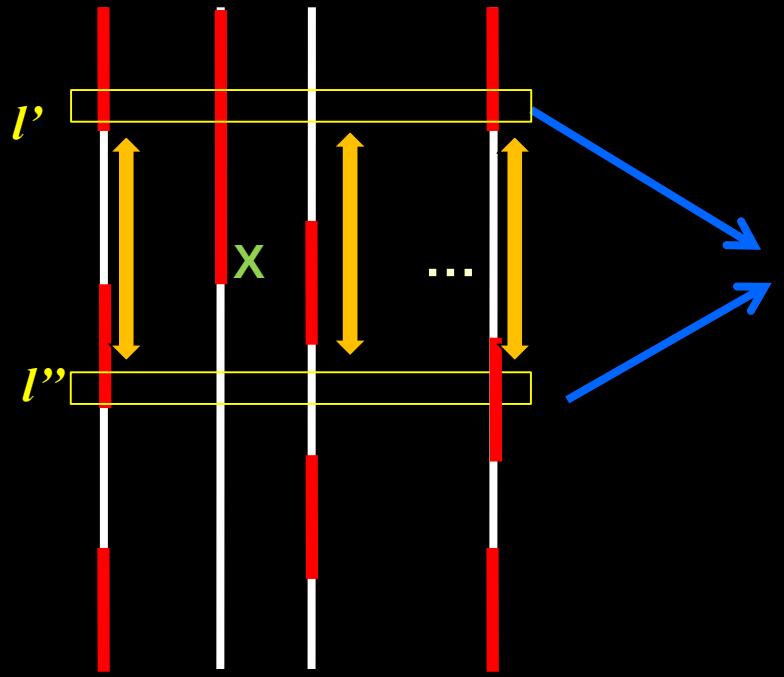
Strain 0

Strain 1

R Resistant

R Resistant

Progeny strains (genotype)



Progeny strains (phenotype)

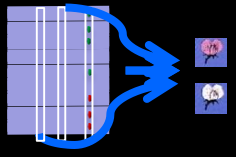
R N R … R

Find pair of loci *l',l''*

0,0 ⟶ R

1,1 ⟶ R

0,1 ⟶ NR



Locus Compatibility approach

Parental strains

Strain 0

Strain 1

R **Resistant**

R **Resistant**

Progeny strains (genotype)

l'

X

…

l''

Progeny strains (phenotype)

R N R … R

!

Find pair of loci *l',l''*

0,0 ⟶ R

1,1 ⟶ R

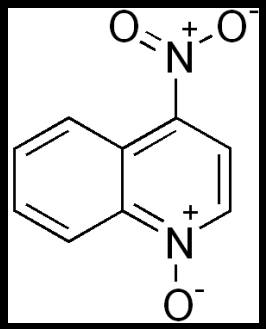
0,1 ⟶ NR

Computational challenge – limiting the number of pairs tested

Application to yeast DNA repair phenotype

S. cerevisiae DNA repair phenotype dataset (Demogines et al, 2008).

response to 4-NQO treatment



Both parents – resistant

Progenies - 31 sensitive; 53 resistant;

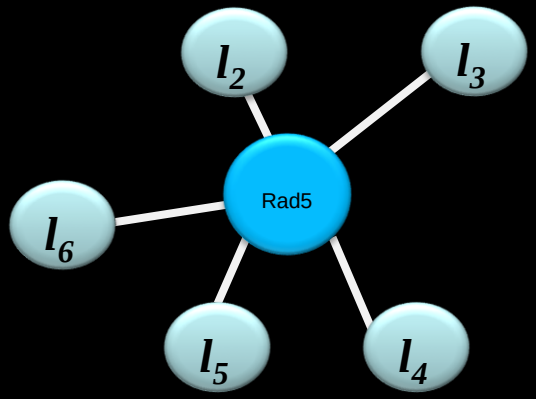
Application to yeast DNA repair phenotype

RESULT:

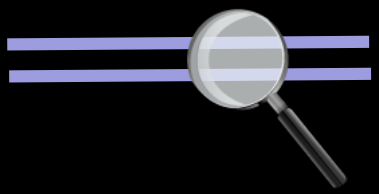
5 candidate interacting pairs of loci

(l_1,l_2) , (l_1,l_3) , (l_1,l_4) , (l_1,l_5) , (l_1,l_6)

l_1 locus contains gene **Rad5 –** involved DSB repair processes

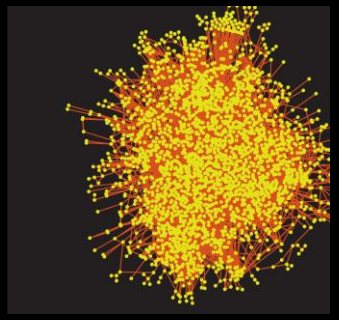


Experimentally confirmed by Trey Ideker and collaborators.



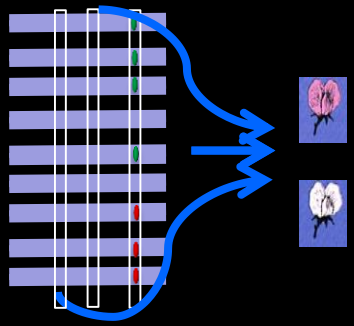
Impact of variations in DNA/RNA sequence and structure on gene expression and function

- Combined experimental and computational evidence for occurrences of non B-DNA structures in *vivo*
- Method to measure impact of SNPs of mRNA structure



Propagation of genetic perturbations through the molecular system

- Method to identify dys-regulated pathways in complex diseases
- Identification of causal genes



Epistatic interactions

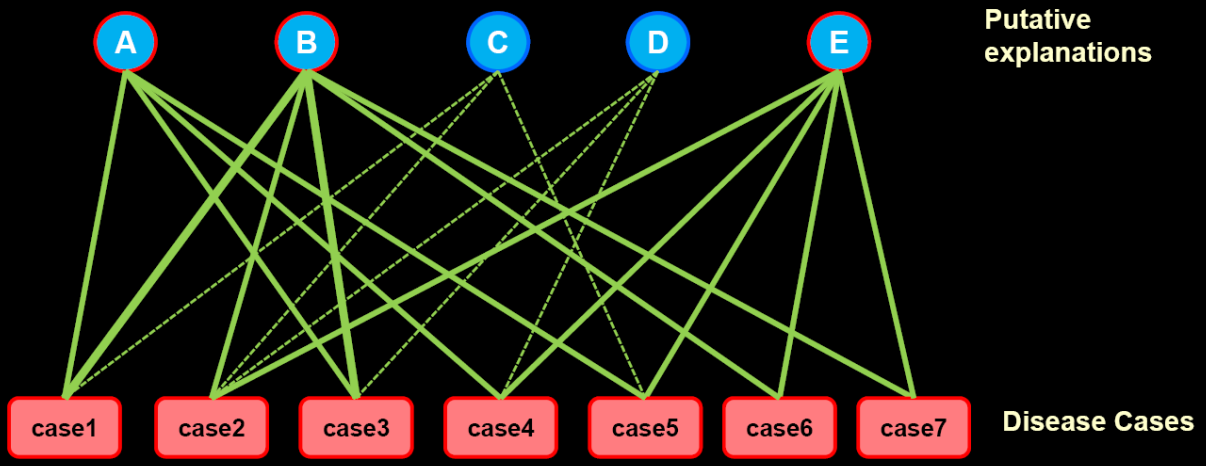
- A new method to study drug resistance

Acknowledgments

Group members:	Topic/interests	Collaborators:
DongYeon Cho	Dros_del Machine learning	Brian Oliver (NIDDK): John Malone, Nicolas Mattiuzzo
Yang Huang	Epistasis / eQTL graph theory	Justin Andrews (Indiana University) Trey Ideker (UCSD): Rohith Srivas
Yoo-Ah Kim	cancer networks optimization	Stefan Wuchty (NCBI) Jozef Przytycki (GWU)
Raheleh Salari	RNA structure Combinatorial algorithms	Michael Gottesman (CCR): Andrew Fung Chava Kimchi-Sarfaty (FDA)
Damian Wojtowicz	Non-B-DNA Evolution Markov Chains	David Levens (NCI): Fedor Kouzine Rafael Casellas (NIAMS): Arito Yamane, Wolfgang Resch
		Craig Benham (UC, Davis)

Selecting representative set of target genes using set cover approach

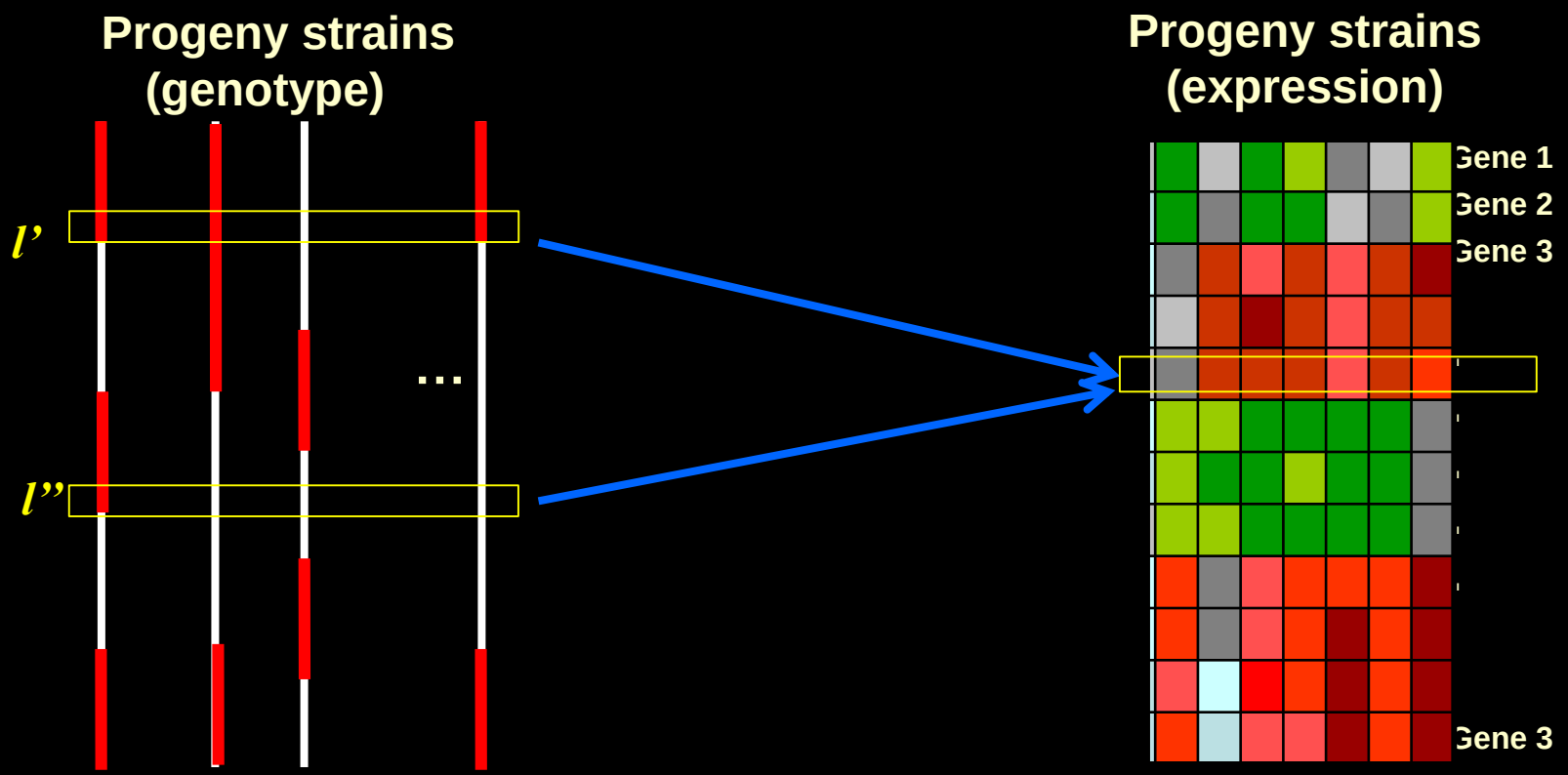
Find smallest set of genes that explain (almost)
all cases such that each case is explained k times



Epistatic model for effects on gene expression

independent model $y = b_0 + b_1 x' + b_2 x'' + \varepsilon$

Epistatic model $y = b_0 + b_1 x' + b_2 x'' + b_3 x' x'' + \varepsilon$



Huang *et al* submitted

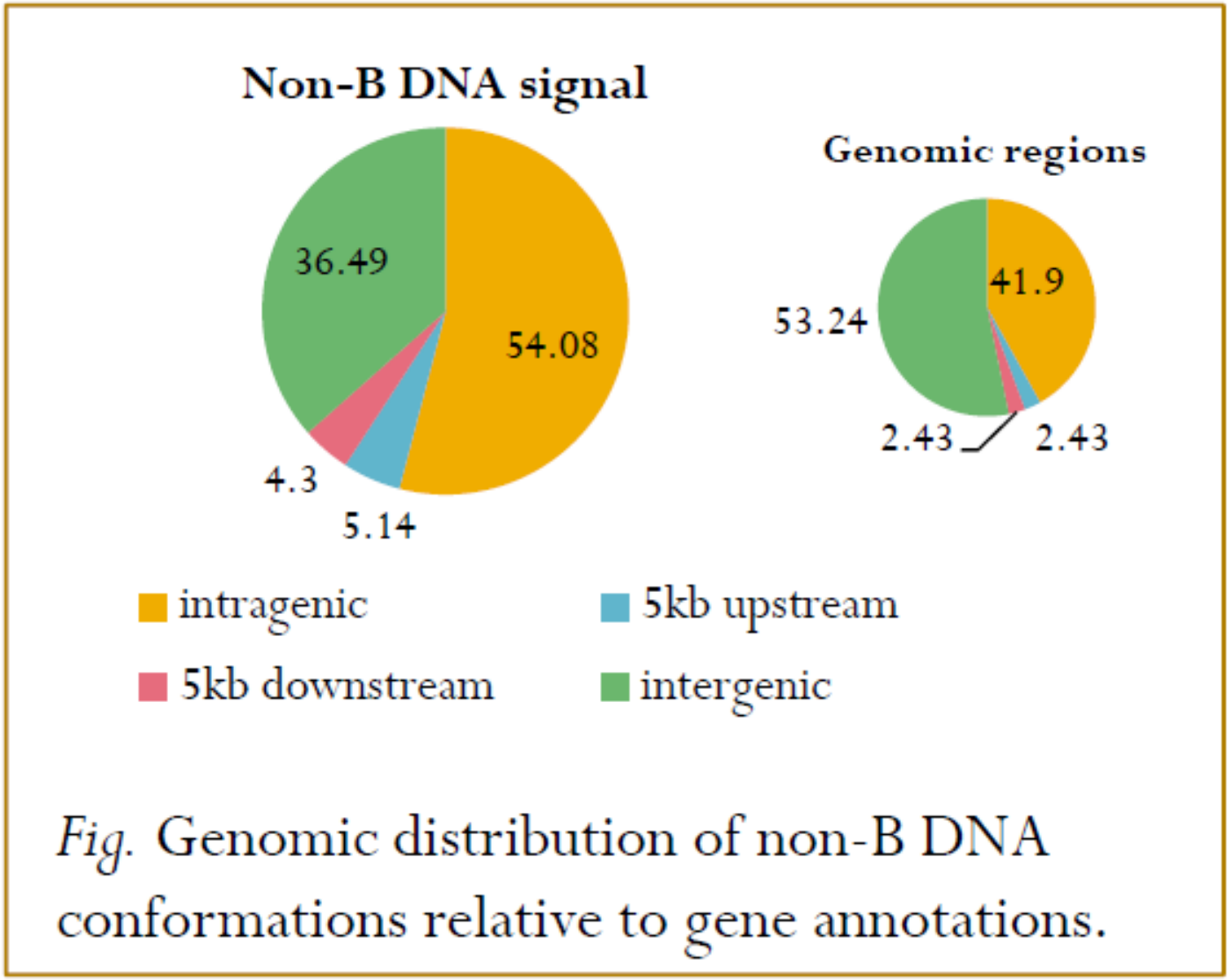
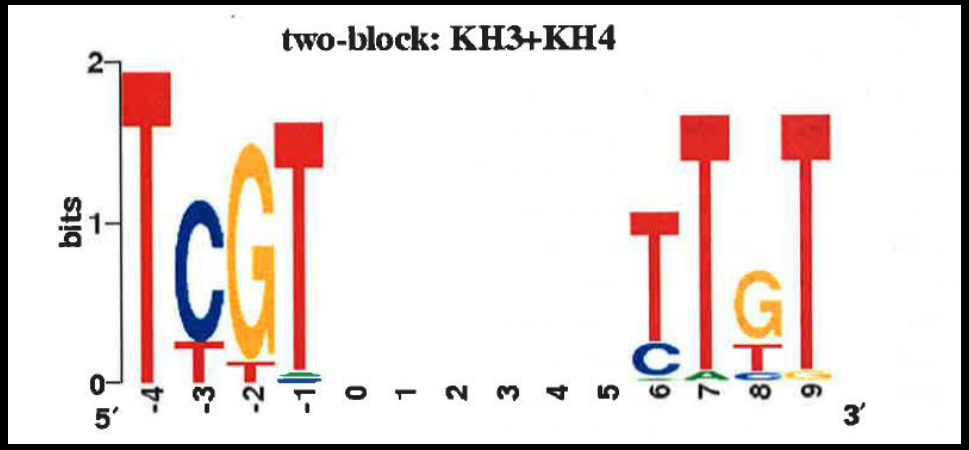
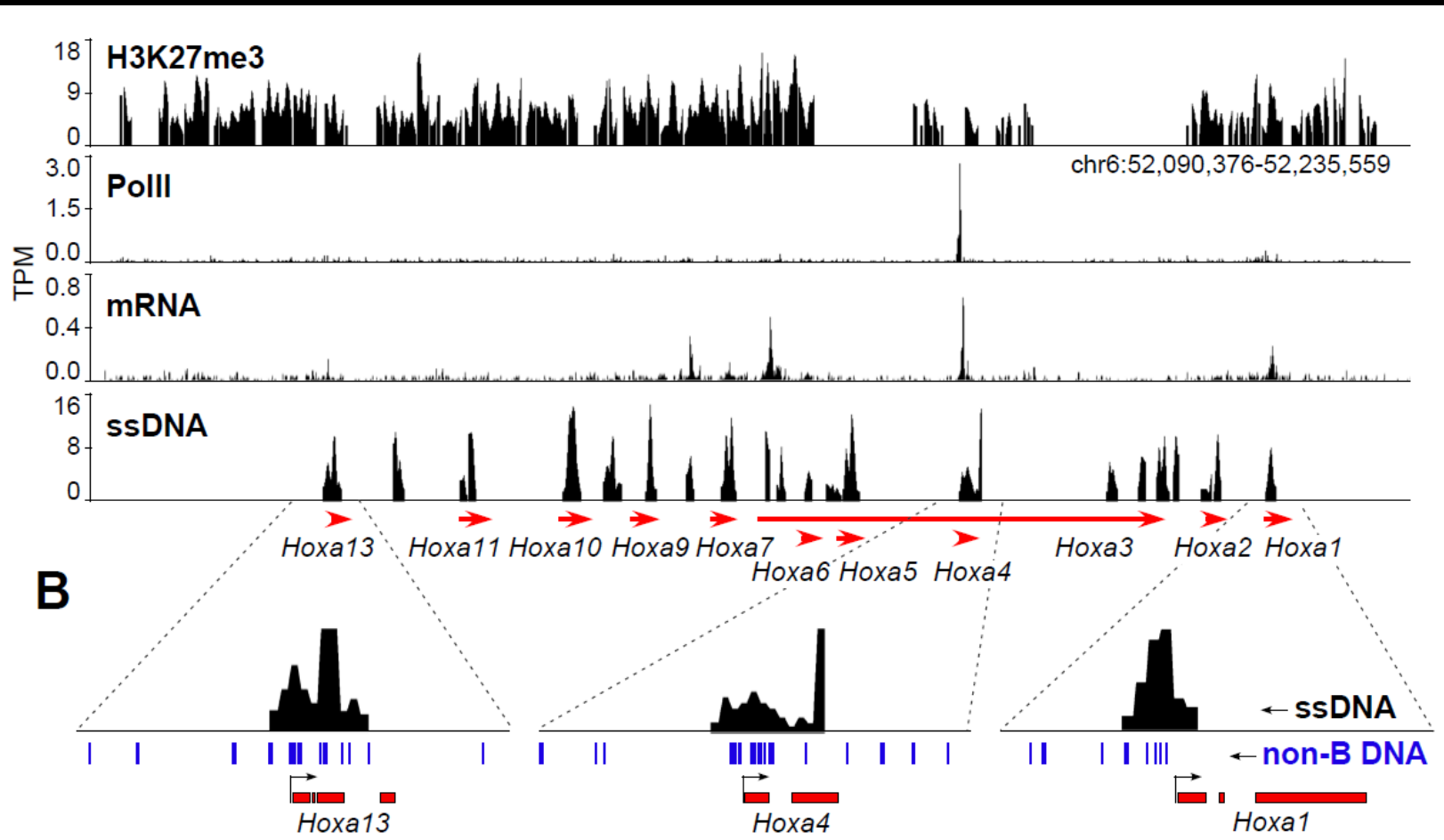


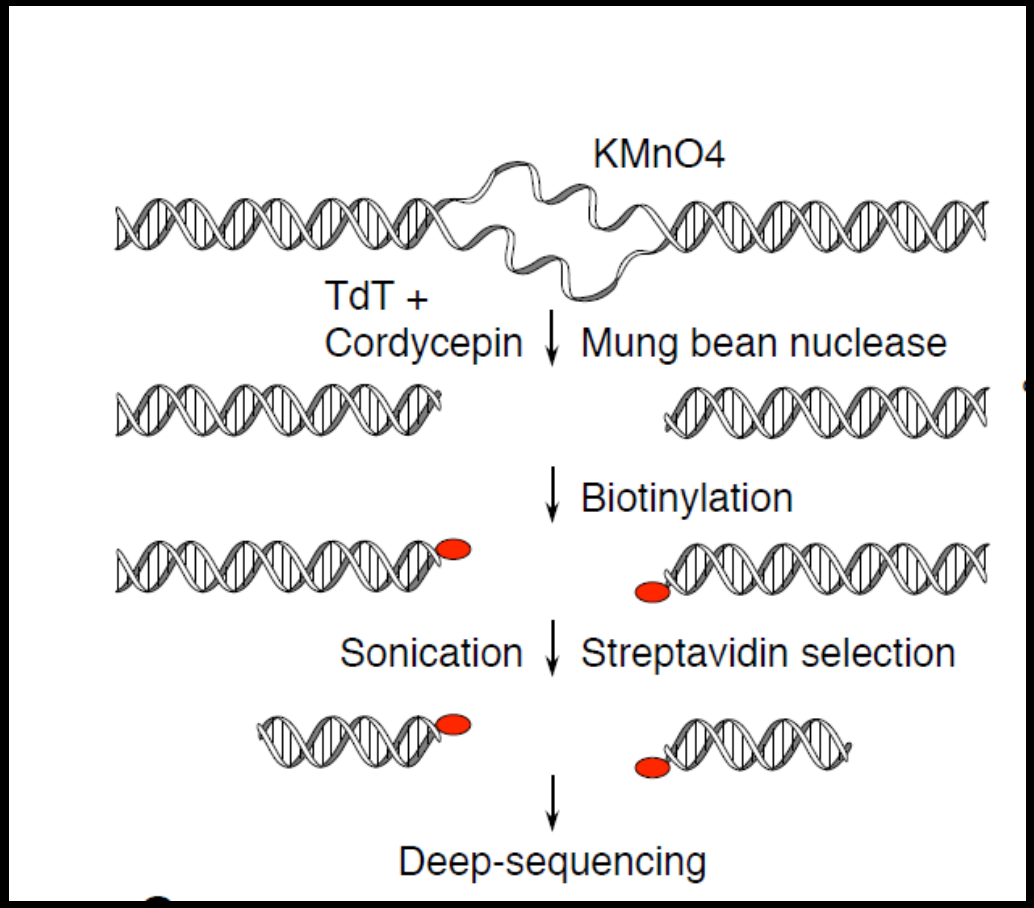
Fig. Genomic distribution of non-B DNA conformations relative to gene annotations.

- Regulatory regions
- Mutation in coding regions
- Copy number variations (CNV)



Binding motifs for FBP domains KH3+HK4 inferred based on paper by Benjamin, Chung, Sanford, Kouzine, Liu, Levens, *PNAS 2006*

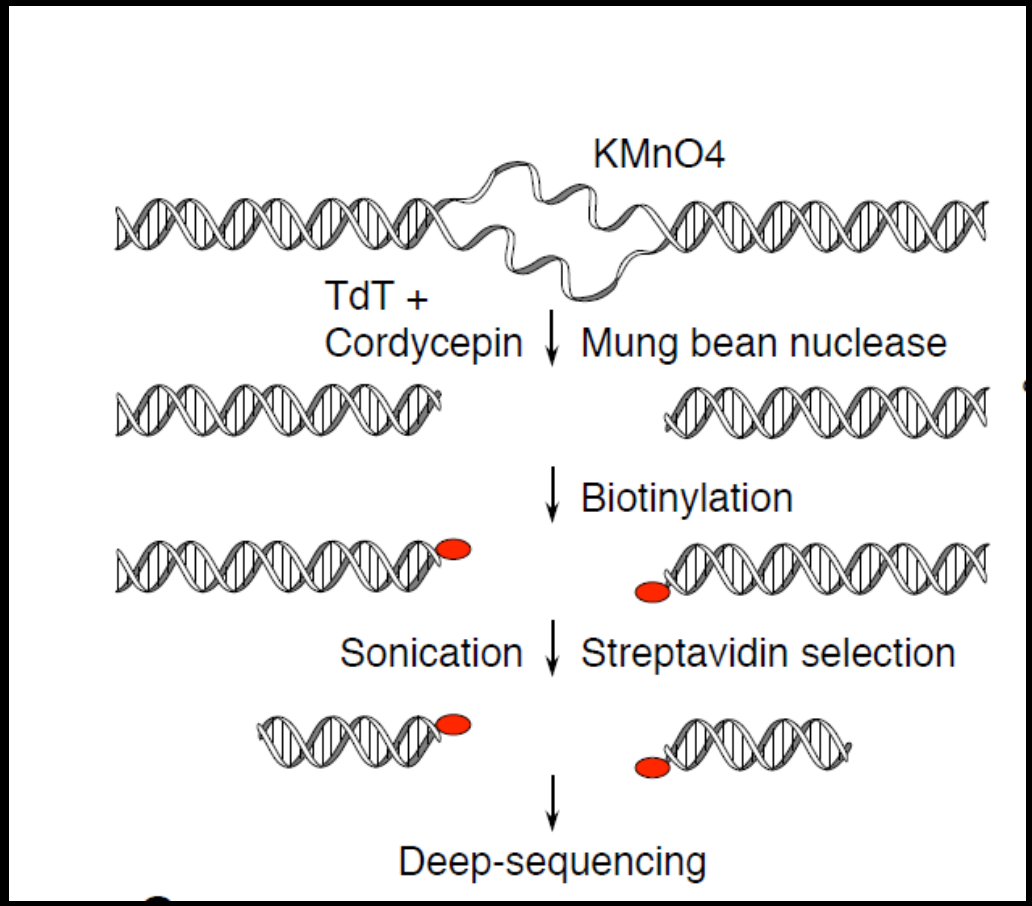

H3K27me3
18
9
0
PolII
3.0
1.5
0.0
chr6:52,090,376-52,235,559
TPM
mRNA
0.8
0.4
0.0
ssDNA
16
8
0
Hoxa13
Hoxa11
Hoxa10
Hoxa9
Hoxa7
Hoxa6
Hoxa5
Hoxa4
Hoxa3
Hoxa2
Hoxa1
B
ssDNA
non-B DNA
Hoxa13
Hoxa4
Hoxa1



- Properties related to promoter melting

  Kouzine, Wojtowicz, Yamada et al. in preparation

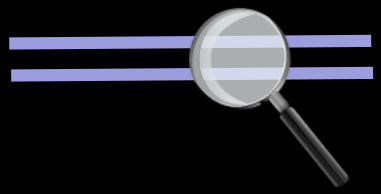
- In vivo detection of Non-B-DNA structures



- Properties related to promoter melting

  Kouzine, Wojtowicz, Yamada et al. in preparation

- In vivo detection of Non-B-DNA structures



Impact of silent mutations

- Codon usage
- mRNA structure

- Codon use optimized in evolutions for translation efficiency

- Codon use is associated with expression noise

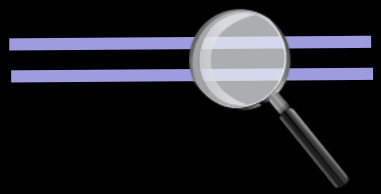
  Salari, Zheng, Wojtowicz-manuscript in preparation

- Codon use is optimized to reduce translation errors including frame-shifting

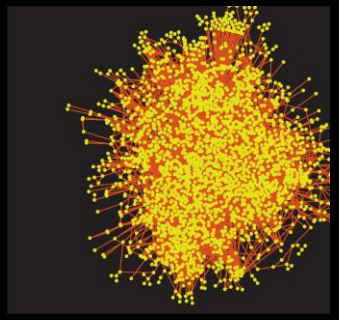
  Huang, Koonin, Lipman, Przytycka NAR 2007

- By altering translation speed can impact the dynamics of protein folding

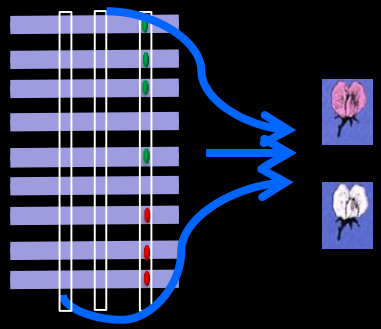
  - Codon dependent speed of translation
  - mRNA structure



Impact of variations in DNA/RNA sequence structure on gene function



Propagation of genetic perturbations through the molecular system



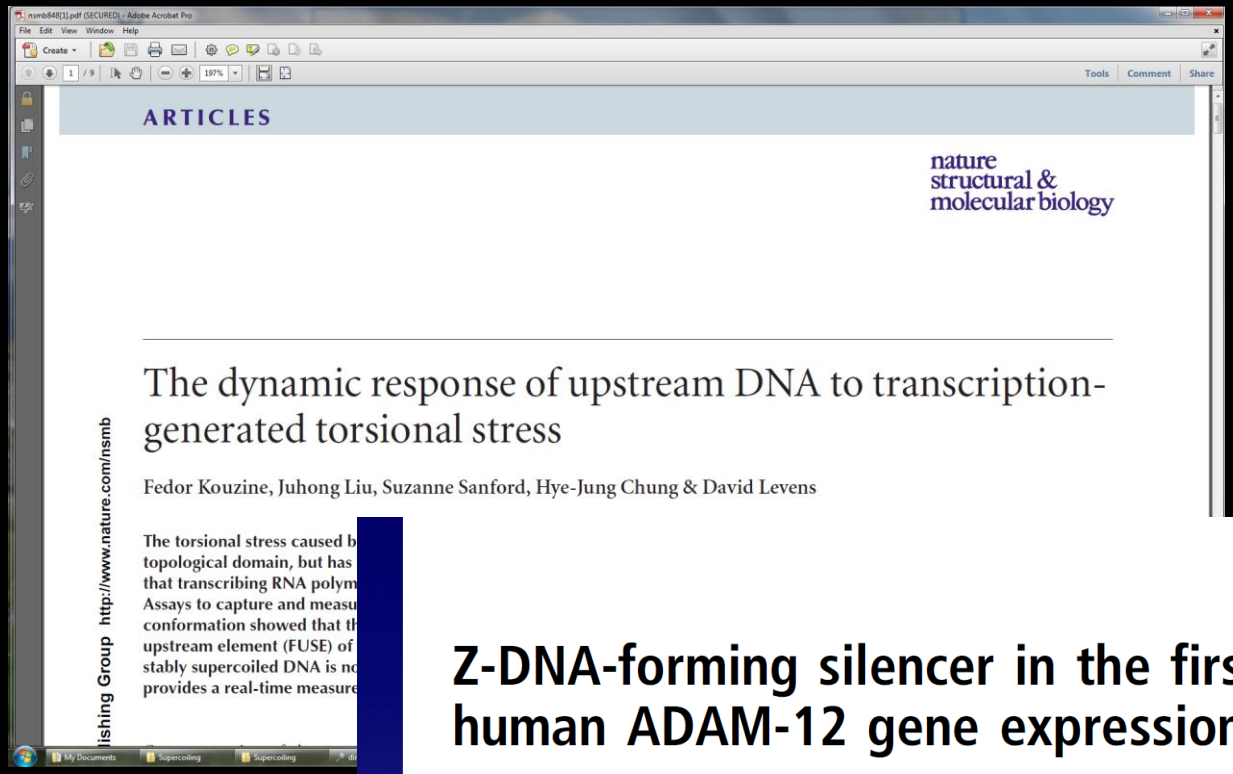
Uncovering epsitatic interactions

Regulatory role of DNA conformational changes in DNA structure

collaboration with *David Levens* (NCI) and *Rafael Casellas* (NIAMS)

Special acknowledgments:

Damian Wojtowicz (TP), *Fedor Kouzine* (DL), *Arito Yamada* (RC)



ARTICLES

nature structural & molecular biology

The dynamic response of upstream DNA to transcription-generated torsional stress

Fedor Kouzine, Juhong Liu, Suzanne Sanford, Hye-Jung Chung & David Levens

The torsional stress caused b
topological domain, but has
that transcribing RNA polym
Assays to capture and measu
conformation showed that th
upstream element (FUSE) of
stably supercoiled DNA is nc
provides a real-time measure

ishing Group http://www.nature.com/nsmb

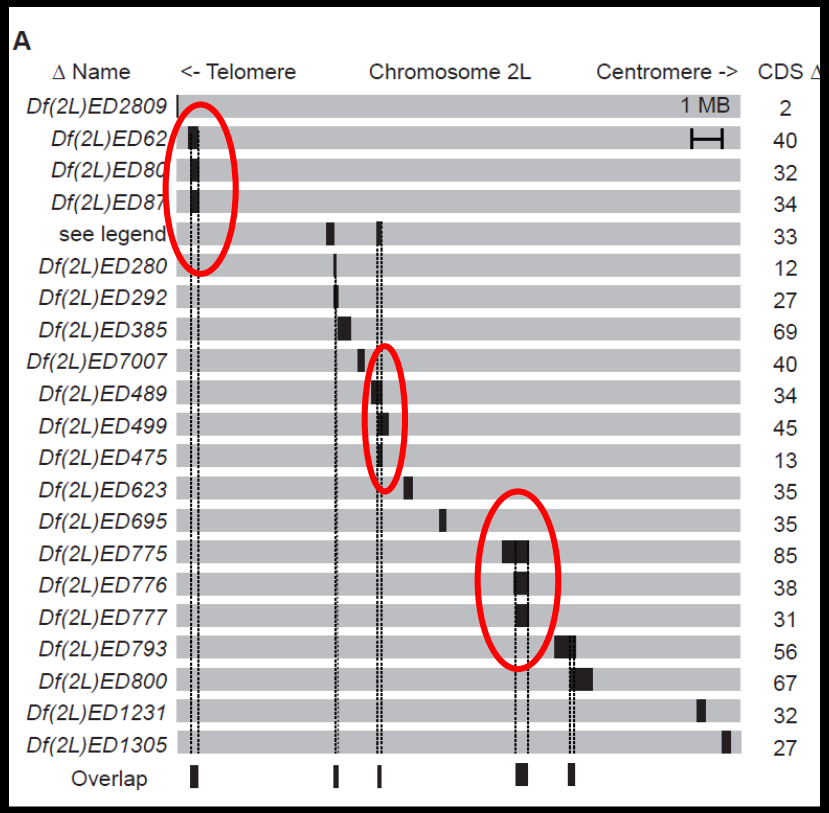
Z-DNA-forming silencer in the first exon regulates human ADAM-12 gene expression

Bimal K. Ray, Srijita Dhar, Arvind Shakya, and Alpana Ray[1]

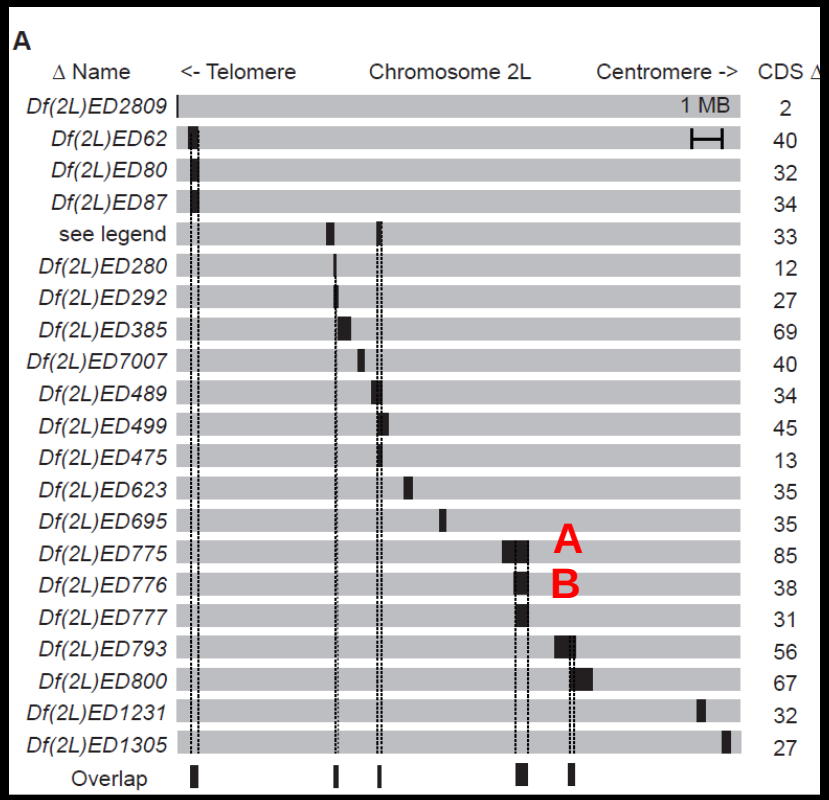
Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211

Edited* by Alexander Rich, Massachusetts Institute of Technology, Cambridge, MA, and approved November 19, 2010 (received for review June 21, 2010)

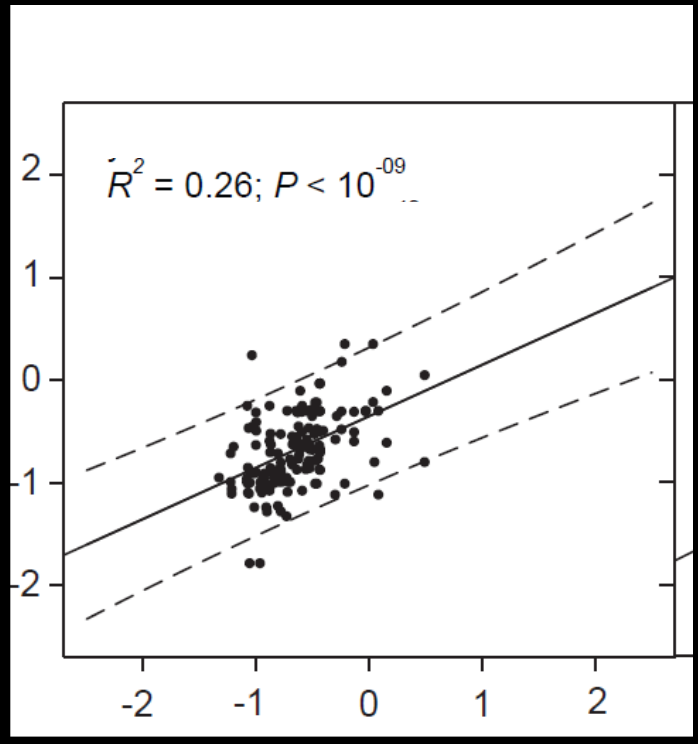
Is response to CNV gene specific?



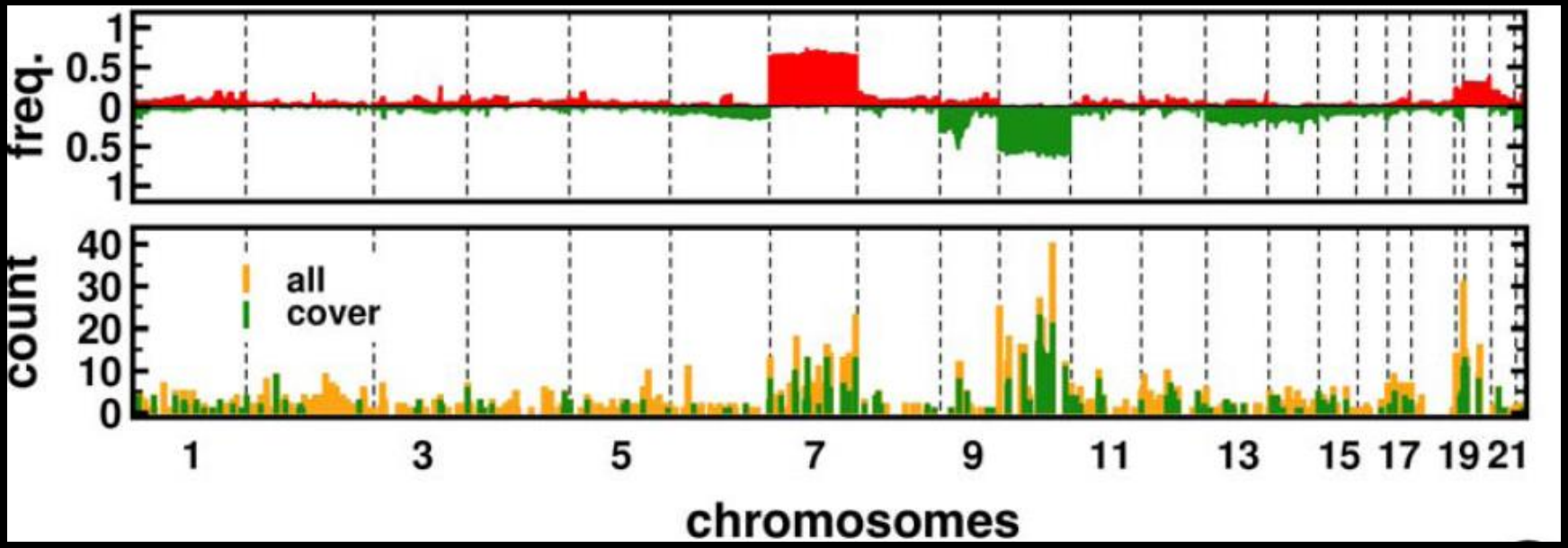
Is response to CNV gene specific?



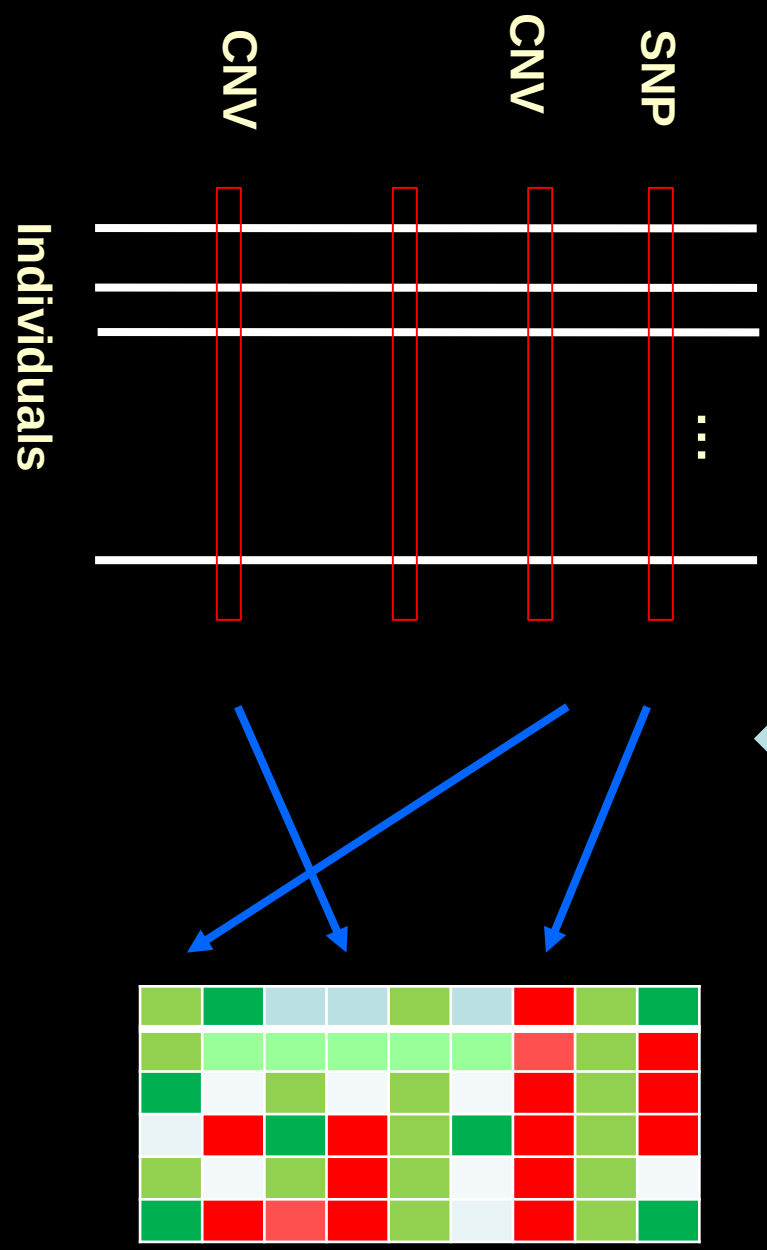
$$\log_2 \frac{Expr(g, DefB)}{Expr(g, WildType)}$$

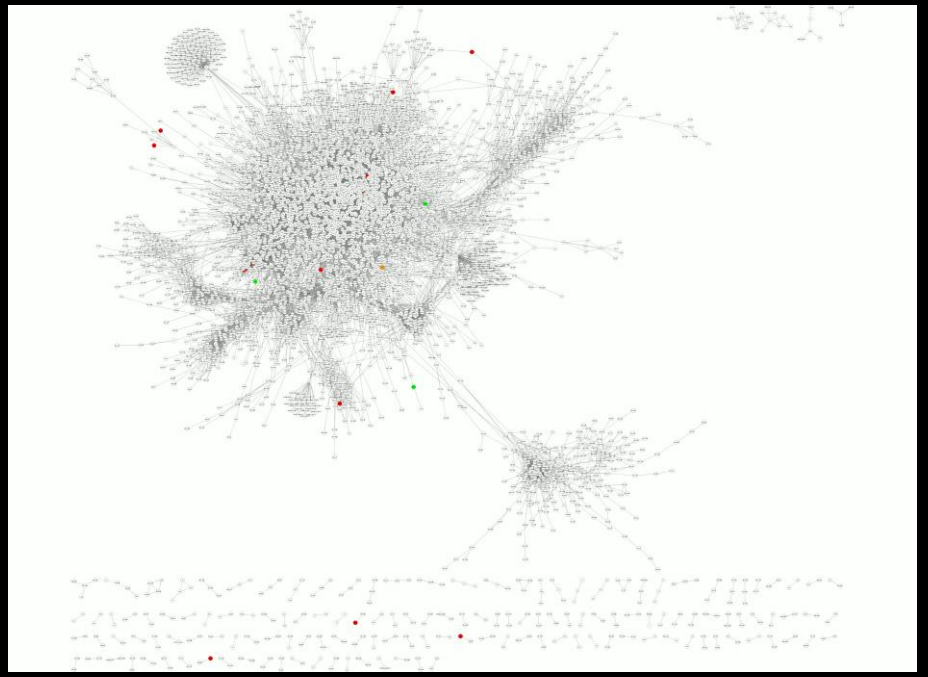


$$\log_2 \frac{Expr(g, DefA)}{Expr(g, WildType)}$$



Starting from selecting “disease genes” we identified copy number variations that associate with expression changes of these genes and putative pathways that propagate the genetic perturbation from copy number variation to the disease genes





Phenotypes

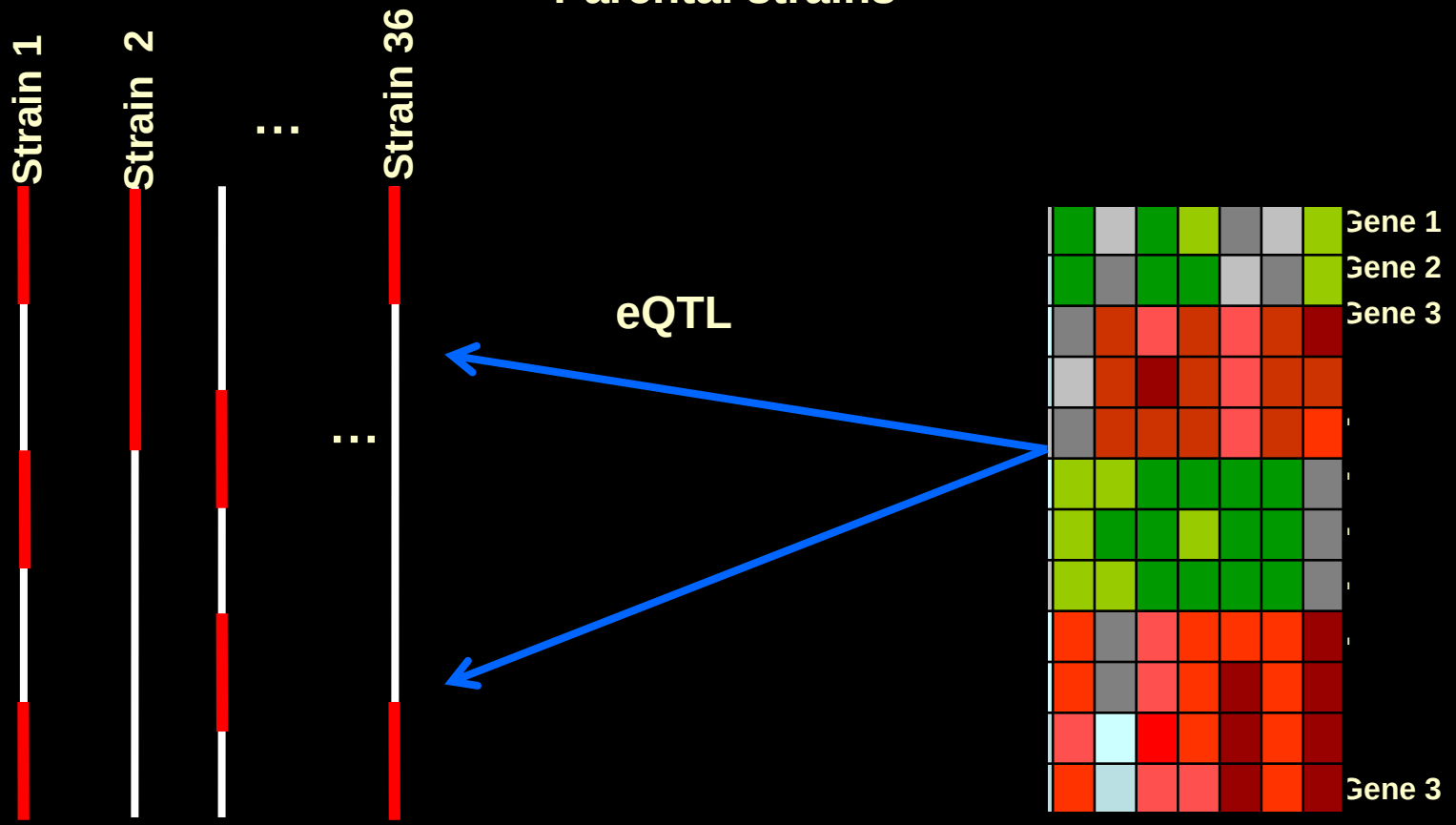
independent model

$$y_i = b_0 + b_1 x_{j1} + b_2 x_{j2} + \varepsilon_i$$

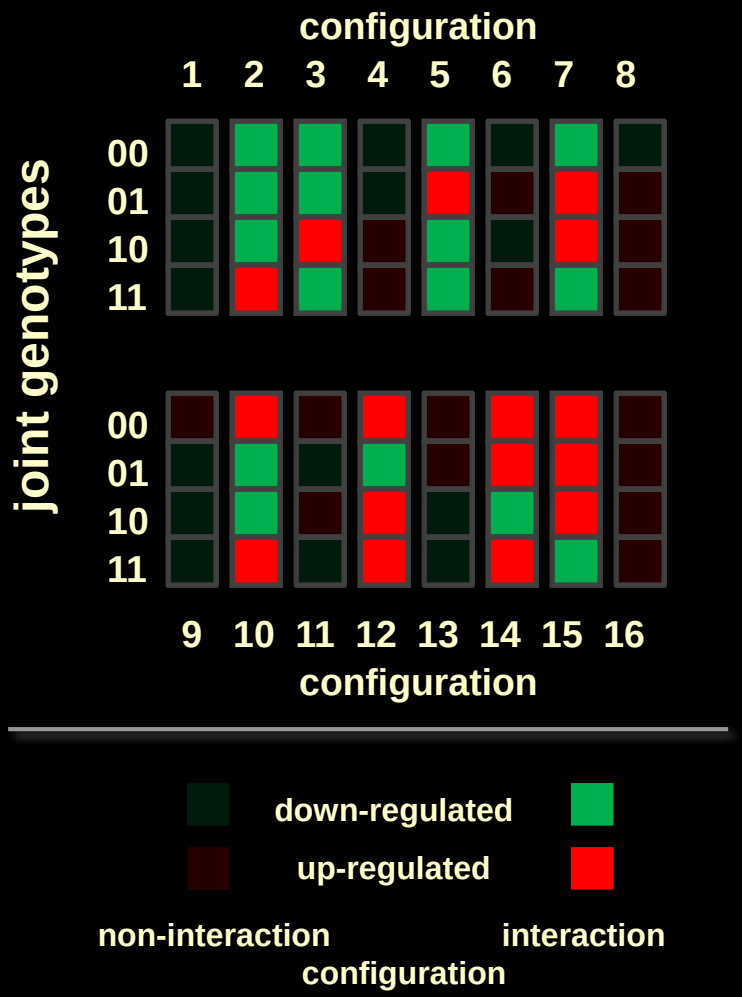
Epistatic model

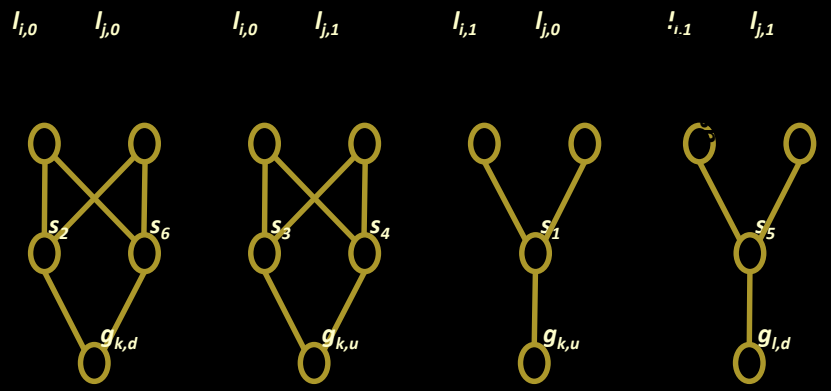
$$y_i = b_0 + b_1 x_{j1} + b_2 x_{j2} + b_3 x_{j1} x_{j2} + \varepsilon_i$$

Parental strains

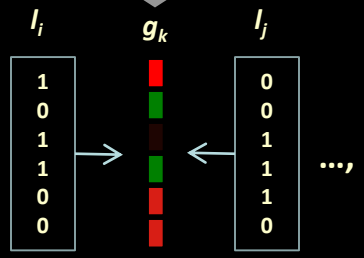


Yang Huang





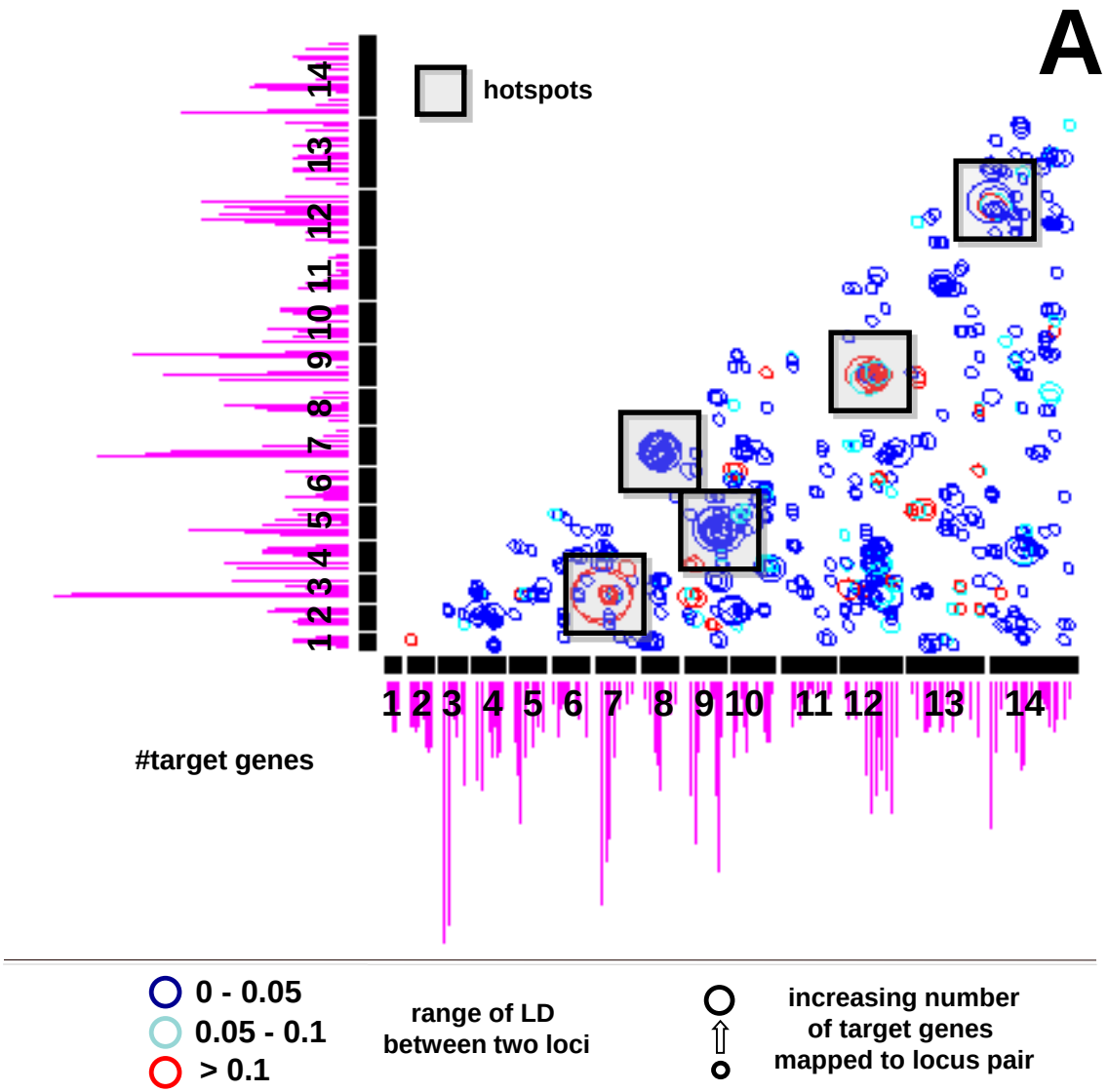
candidate sets:
… , (l_i, l_j, g_k), …

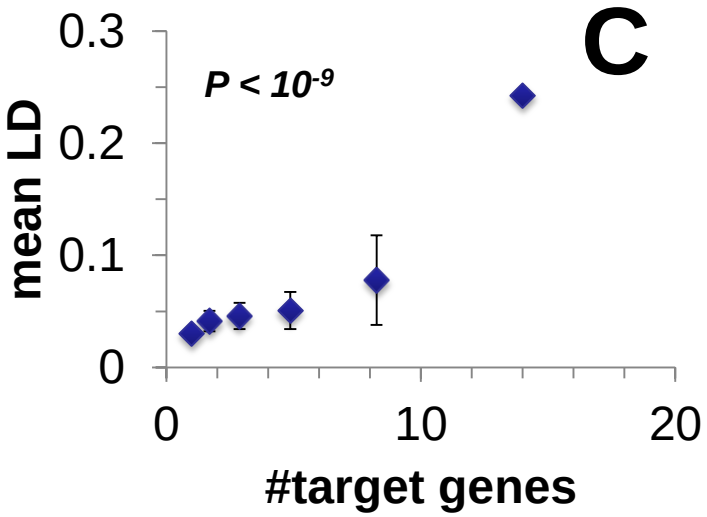


epistatic interactions

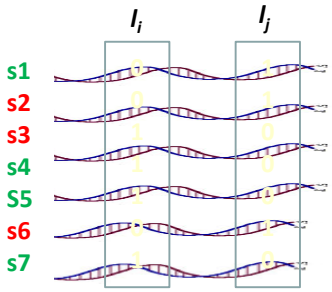
Huang et al. submitted

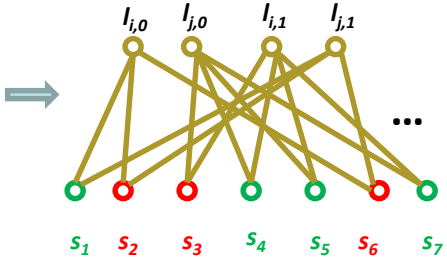
Yang Huang


A
hotspots
1 2 3 4 5 6 7 8 9 10 11 12 13 14
#target genes
0 - 0.05
0.05 - 0.1
> 0.1
range of LD
between two loci
increasing number
of target genes
mapped to locus pair


C
P < 10-9
mean LD
0.3
0.2
0.1
0
0
10
20
#target genes

Yang Huang

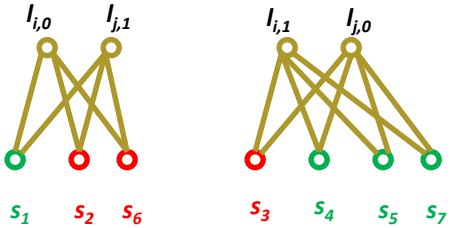




Build QTL association graph

loci

progenies



Identify phenotype inclined cliques

$(l_{i,0}, l_{j,1})$ in a mostly resistant clique

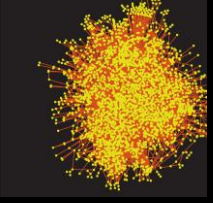
$(l_{i,1}, l_{j,0})$ in a mostly sensitive clique

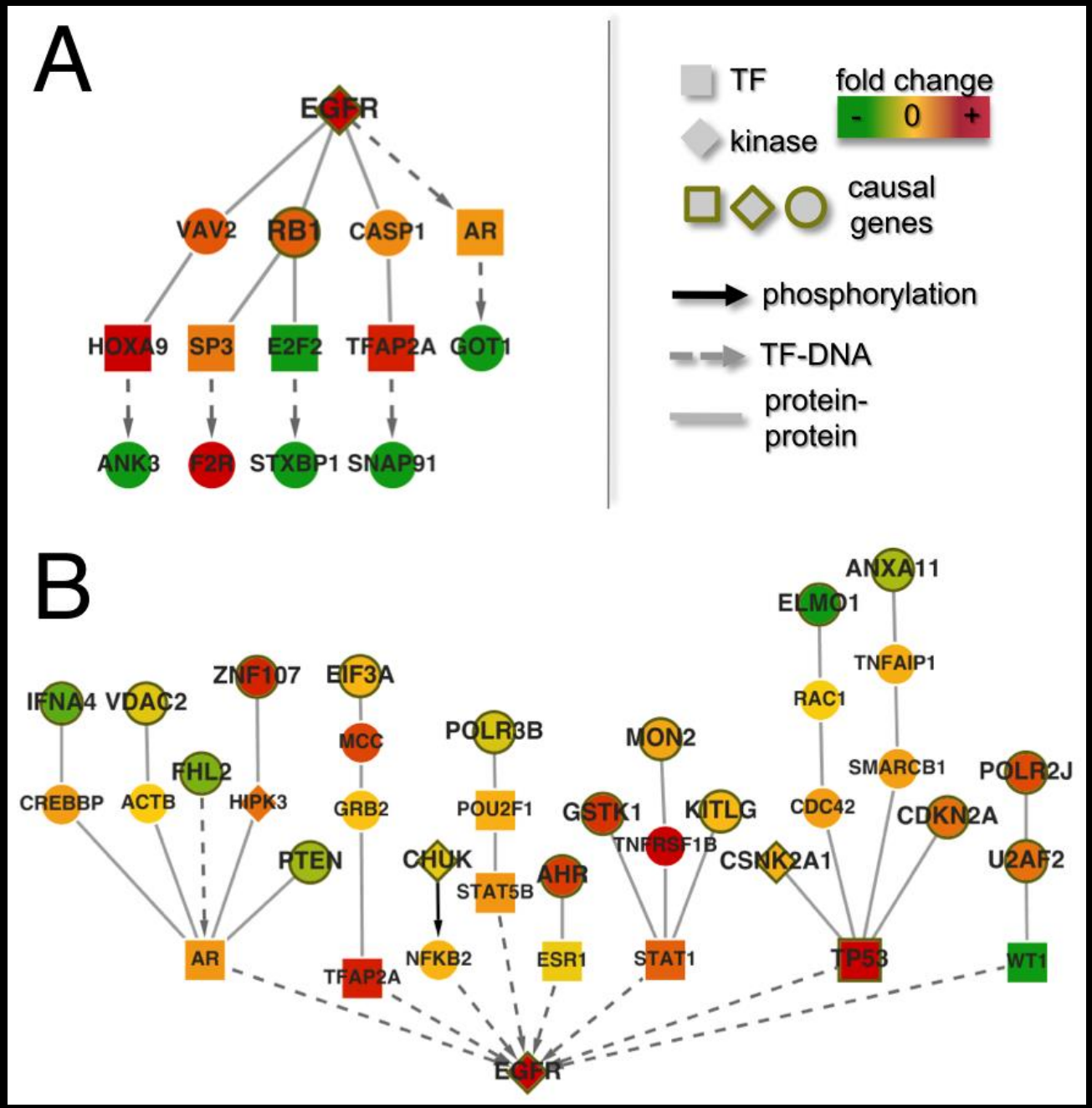
candidate interaction locus pairs
…, (l_i, l_j), ….

determine Interacting loci

progenies with sensitive phenotype

progenies with resistant phenotype

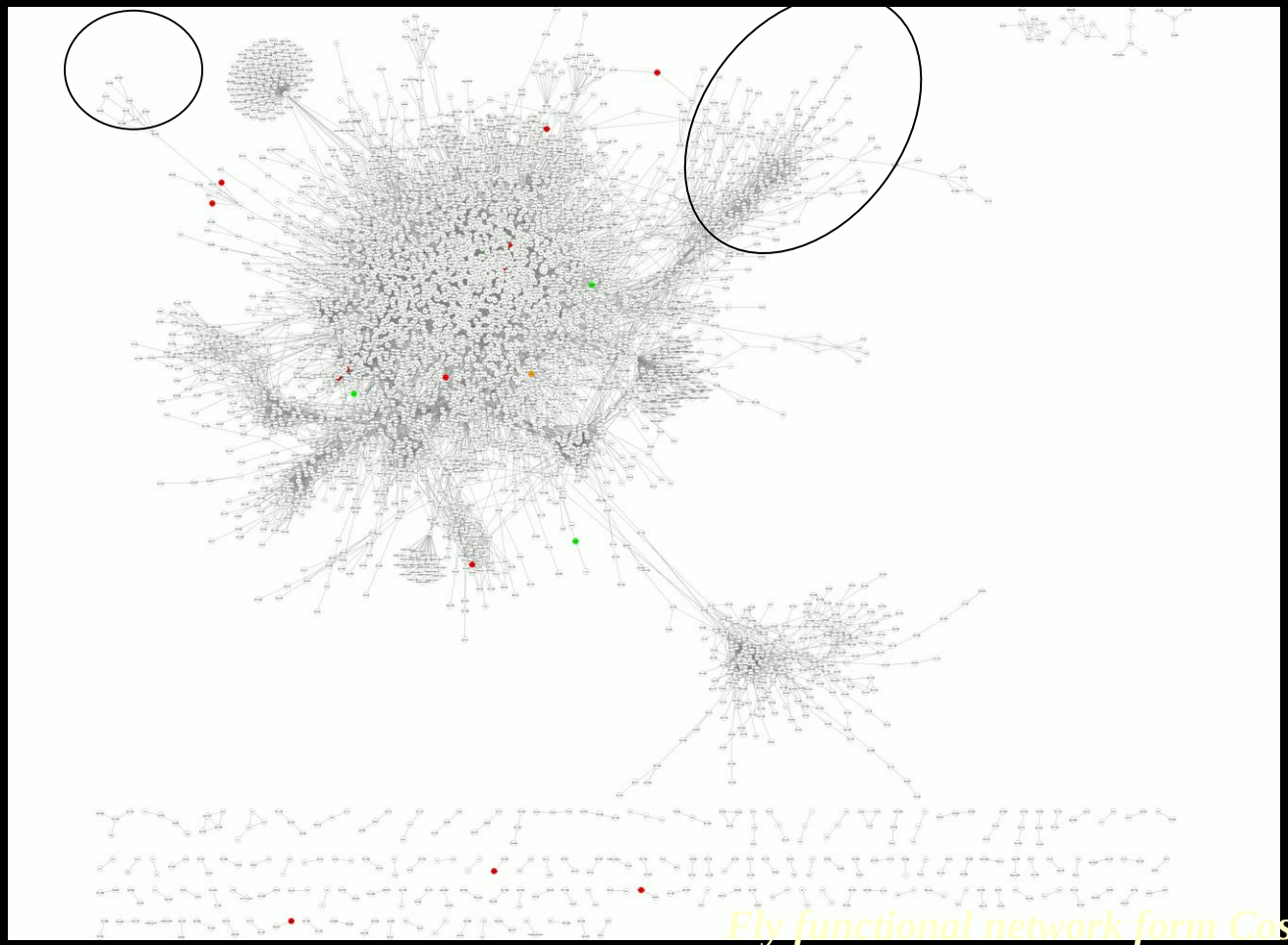



A
EGFR
VAV2
RB1
CASP1
AR
HOXA9
SP3
E2F2
TFAP2A
GOT1
ANK3
F2R
STXBP1
SNAP91
TF
kinase
causal genes
fold change
-
0
+
phosphorylation
TF-DNA
protein-protein
B
ANXA11
ELMO1
TNFAIP1
RAC1
SMARCB1
POLR2J
CDC42
CDKN2A
U2AF2
CSNK2A1
TP53
WT1
ZNF107
EIF3A
IFNA4
VDAC2
MCC
POLR3B
MON2
FHL2
CREBBP
ACTB
HIPK3
GRB2
POU2F1
GSTK1
KITLG
TNFRSF1B
PTEN
CHUK
AHR
STAT5B
AR
TFAP2A
NFKB2
ESR1
STAT1
EGFR

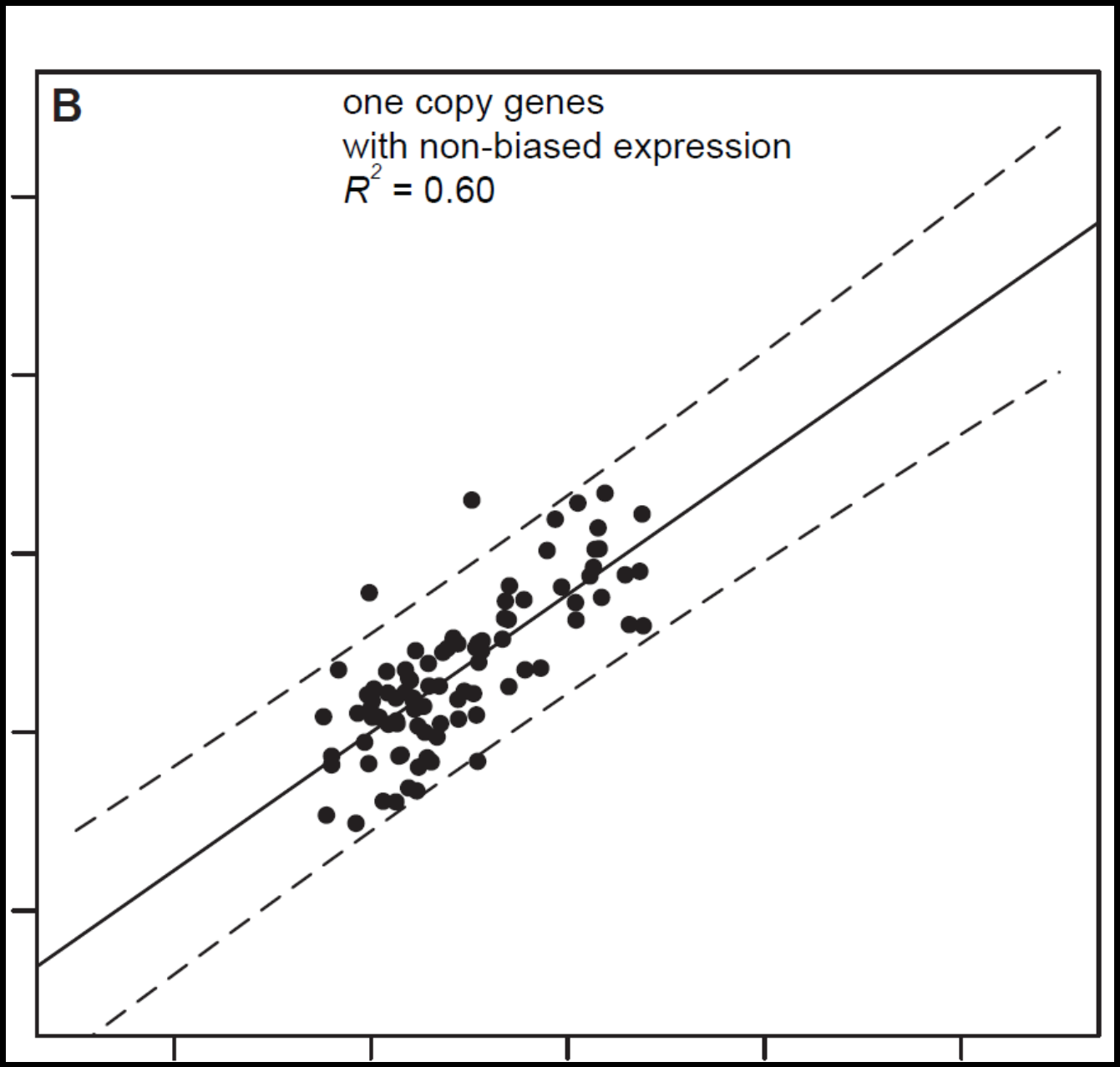
Propagation of the effects of genetic changes (DrosDel)

collaboration with *Brian Oliver* group (NIDDK)

Special acknowledgments:
Jon Malone (BO), *DongYeon Cho* (TP)



Fly functional network form Costello et.al GB 2009

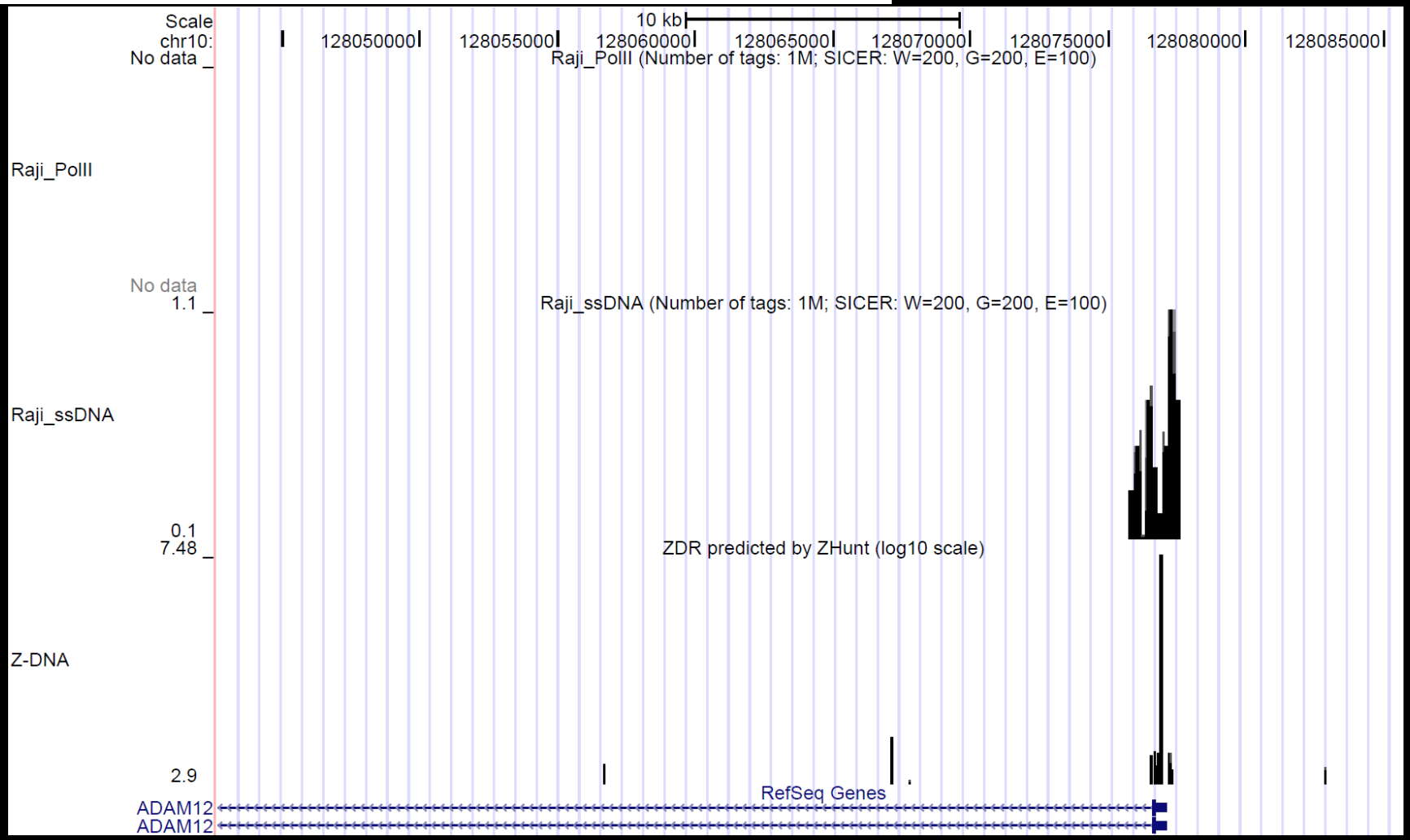
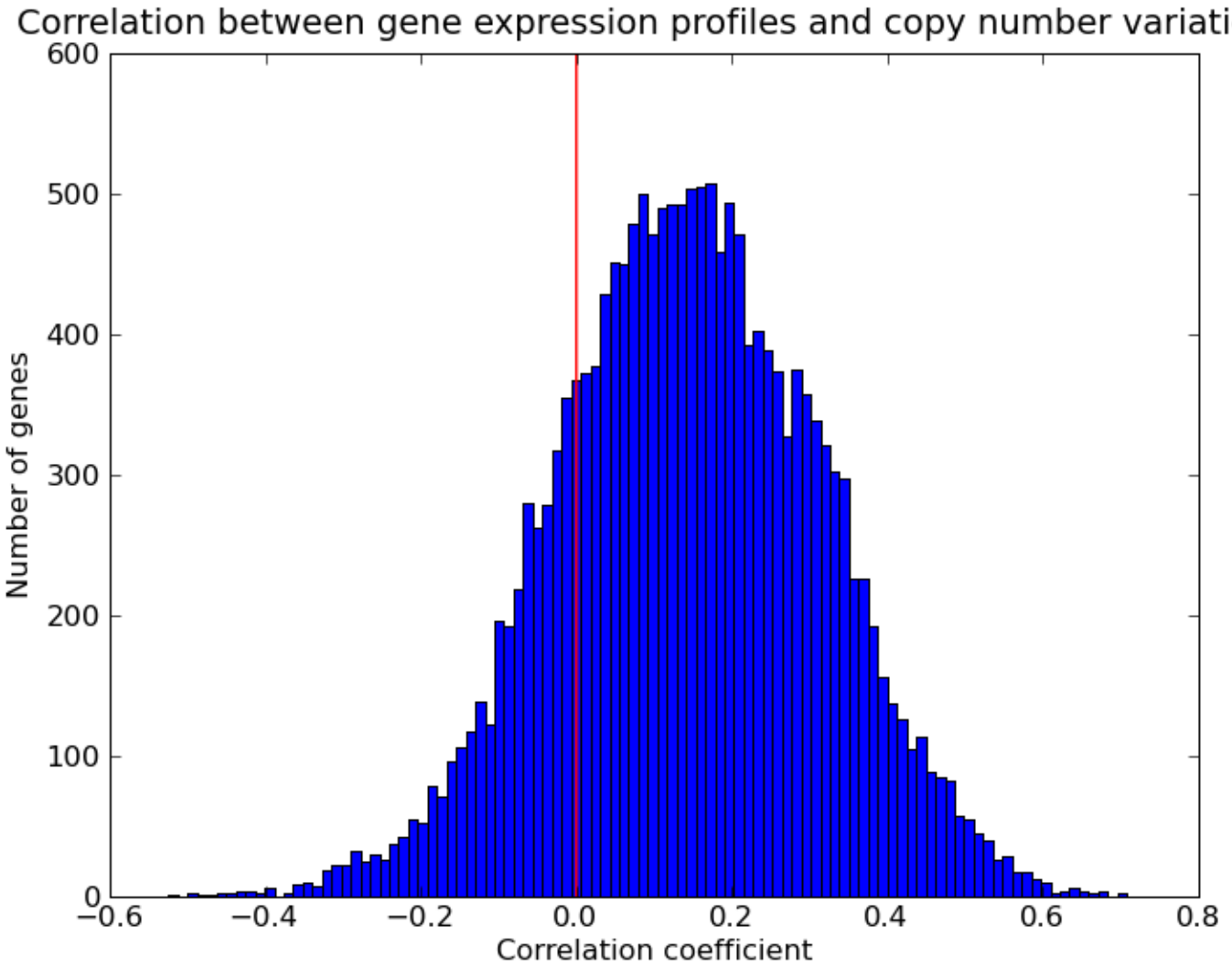

B
one copy genes
with non-biased expression
R^2 = 0.60

Z-DNA-forming silencer in the first exon regulates human ADAM-12 gene expression

Bimal K. Ray, Srijita Dhar, Arvind Shakya, and Alpana Ray[1]

Department of Veterinary Pathobiology, University of Missouri, Columbia, MO 65211

Edited* by Alexander Rich, Massachusetts Institute of Technology, Cambridge, MA, and approved November 19, 2010 (received for review June 21, 2010)



Form our brain cancer studies: